UNIVERSITÉ DU QUÉBEC À MONTRÉAL

APPLICATION DE TECHNIQUES DE FORAGE DE TEXTES DE NATURE PRÉDICTIVE ET EXPLORATOIRE À DES FINS DE GESTION ET D'ANALYSE THÉMATIQUE DE DOCUMENTS TEXTUELS NON STRUCTURÉS

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE COGNITIVE

PAR

DOMINIC FOREST

JUIN 2006

- Tu travailles ?
- J'essaie de travailler : c'est bien plus difficile.

Jules Renard

AVANT-PROPOS

[...] The great academic tradition [is to know] more and more about less and less until you know everything about nothing.

S. Pinker

Cette thèse de doctorat s'inscrit dans le prolongement de notre mémoire de maîtrise en philosophie finalisé en mai 2002. Dans le cadre de ces travaux antérieurs (Forest, 2002; Forest et Meunier, 2000), nous avons exploré la possibilité d'assister informatiquement l'analyse thématique de textes philosophiques. Ces travaux ont mené à la démonstration qu'il était théoriquement possible d'assister l'analyse thématique de textes philosophiques. Par ailleurs, nous avons aussi validé les résultats obtenus automatiquement avec les interprétations classiques du corpus philosophique analysé. D'un point de vue informatique, notre contribution s'est limitée à explorer la possibilité d'appliquer un processus de classification ou de regroupement non supervisé afin de regrouper différents passages de notre corpus d'expérimentation. Sur la base des résultats de la classification, nous avons, par la suite, observé que certains termes présents dans plusieurs regroupements pouvaient s'avérer utiles afin de découvrir les documents traitant d'un sujet ou d'un thème particulier. Dans ces travaux, le processus visant à identifier le contenu des regroupements a été réalisé manuellement.

Le projet de recherche doctoral que nous proposons traite de la problématique de la gestion et de l'analyse thématique des documents textuels non structurés. Notre projet repose donc sur les acquis et les conclusions de nos travaux antérieurs. Il vise à poursuivre le travail entrepris durant les dernières années en approfondissant certaines réflexions entamées antérieurement, en explorant d'autres techniques de traitement et d'analyse et en abordant la

problématique de l'analyse thématique non pas du point de vue de l'analyse de textes philosophiques, mais plutôt dans une perspective, plus proche des préoccupations des sciences de l'information, d'analyse et de gestion informatisées des documents textuels.

Cette thèse de doctorat s'insère dans le programme multidisciplinaire de doctorat en informatique cognitive de l'Université du Québec à Montréal. La problématique de l'analyse thématique est donc principalement abordée d'un point de vue informatique. Cependant, les différentes réflexions exposées font appel à des concepts et des techniques provenant de plusieurs domaines, parmi lesquels figurent non seulement les sciences cognitives et les sciences de l'information, mais aussi la linguistique et le Traitement Automatique du Langage (TAL). Il s'agit donc d'une recherche de nature pluridisciplinaire dont l'objectif général n'est pas de proposer une théorie ou un algorithme particulier, mais plutôt, comme ce fut le cas lors de l'ensemble de nos travaux antérieurs, de mettre en relation un certain nombre d'hypothèses théoriques, d'explorer plusieurs concepts et d'appliquer des techniques provenant de différentes disciplines afin de proposer une solution aux problèmes complexes de l'identification des thèmes et de l'analyse thématique des documents textuels non structurés.

Ce travail, tout comme ceux que nous espérons poursuivre, se veut une ouverture sur plusieurs disciplines. Comme en témoigne la citation de Pinker en exorde, nous constatons, à regret (quelques initiatives visent à promouvoir les recherches interdisciplinaires, mais il ne s'agit pas d'une tendance généralisée), que plusieurs projets de recherche académiques semblent être menés sans tenir compte des principaux concepts et des techniques provenant de domaines périphériques. Nous croyons que le développement de la science ne peut que bénéficier d'un décloisonnement des disciplines. Modestement, nous espérons que notre travail contribuera à atteindre cet objectif.

* * *

Ce projet n'aurait pu être réalisé sans la précieuse collaboration et le constant soutien de plusieurs professeurs, collègues, parents et amis. Je remercie donc tous ceux et celles qui, durant les dernières années, ont contribué à la réussite de ce projet.

Je tiens plus particulièrement à remercier chaleureusement mon directeur de recherche, Monsieur Jean-Guy Meunier, professeur de philosophie du langage et de sciences cognitives

au Département de philosophie de l'Université du Québec à Montréal et co-directeur (fondateur) du Laboratoire d'ANalyse Cognitive de l'Information (LANCI). Jean-Guy Meunier a su me transmettre son insatiable curiosité, sa passion pour le travail intellectuel rigoureux, ainsi que pour la recherche universitaire, en plus de m'avoir fait découvrir le vaste domaine de la Lecture et de l'Analyse de Textes Assistées par Ordinateur (LATAO). Plusieurs des idées présentées dans ce mémoire s'inspirent de celles développées, depuis plus de vingt-cinq ans, par Jean-Guy Meunier. Je le remercie vivement aussi pour m'avoir permis de bénéficier du meilleur environnement de recherche que puisse imaginer un étudiant de doctorat.

J'ai connu Jean-Guy Meunier il y déjà dix ans dans son cours d'introduction à la philosophie du langage, alors que j'étais étudiant au baccalauréat en philosophie. Il est mon directeur de recherche depuis septembre 1999. Il est, je ne sais exactement depuis quand, mon ami.

Je tiens aussi à remercier mon co-directeur, Monsieur Hakim Lounis, professeur d'informatique cognitive, d'intelligence artificielle et de génie logiciel au Département d'informatique de l'Université du Québec à Montréal. Mon projet doctoral a été l'occasion de parfaire mes connaissances dans les domaines de l'intelligence artificielle et de l'apprentissage machine. En plus de contribuer à ma formation, Hakim Lounis m'a prodigué de précieux conseils qui ont contribué à la qualité de cette thèse de doctorat.

De plus, je remercie les professeurs qui ont accepté de faire partie du comité d'évaluation de cette thèse : Monsieur Daniel Memmi, professeur d'informatique cognitive et d'intelligence artificielle au Département d'informatique de l'Université du Québec à Montréal ; Monsieur Pierre Poirier, professeur de philosophie de l'esprit et de sciences cognitives au Département de philosophie de l'Université du Québec à Montréal et Monsieur John M. Unsworth, professeur de sciences de l'information et de *humanities computing* et doyen de la *Graduate School of Library and Information Science* de l'*University of Illinois at Urbana-Champaign*.

J'adresse aussi mes remerciements à tous les collègues avec lesquels j'ai eu le plaisir de collaborer au laboratoire LANCI et au programme de doctorat en informatique cognitive. Certaines idées que l'on retrouve dans cette thèse puisent leurs racines dans les échanges et les débats que nous avons eus durant les dernières années. Je suis aussi très redevable à Florian

Ferrand qui, malgré son horaire très chargé, a accepté de corriger la version finale de cette thèse. J'adresse mes plus sincères remerciements à mon collègue Sébastien Hélie qui a généreusement accepté que j'utilise dans mes travaux son implémentation du réseau de neurones ART1. Le milieu académique m'apparaît malheureusement de plus en plus compétitif; je me réjouis néanmoins de constater que de véritables collaborations sont encore possibles.

Cette thèse n'aurait pu être menée à terme sans l'inestimable soutien de ma conjointe Sophie. Bourreau de travail, elle a su me servir de modèle, me permettant ainsi de terminer ce projet sans trop dépasser les délais qui m'étaient impartis. Je la remercie profondément pour son soutien inconditionnel.

Finalement, je reconnais l'aide financière du gouvernement du Canada par l'entremise du programme de bourses de doctorat en recherche du Conseil de Recherche en Sciences Humaines du Canada (CRSH).

Dominic Forest

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Depuis les dix dernières années, on observe une hausse considérable du nombre d'initiatives visant à numériser et à rendre disponible le patrimoine informationnel des organisations et des différentes branches du savoir. Les conséquences découlant de ces initiatives sont importantes et très nombreuses. Elles ont entre autres conduit à l'émergence d'applications permettant différentes opérations complexes d'analyse et de gestion des documents. Malgré la diversité de ces applications, on constate que l'ensemble des disciplines reliées à l'analyse et à la gestion des documents textuels sont axées sur la compréhension et l'informatisation des processus d'identification des contenus thématiques et d'analyse thématique.

Le projet que nous présentons aborde précisément les problématiques de l'identification des thèmes et de l'assistance à l'analyse thématique des documents textuels. L'objectif général du projet est de développer et de valider deux méthodologies informatiques fondées respectivement sur la catégorisation et la classification automatiques permettant d'assister efficacement l'identification des thèmes et, surtout, l'analyse thématique des documents textuels. Il vise ainsi à effectuer un transfert de concepts et de méthodologies provenant, d'une part, des recherches théoriques et pluridisciplinaires portant sur l'analyse thématique et, d'autre part, des recherches appliquées en classification et en catégorisation automatiques des données afin de proposer une méthodologie et un prototype d'application flexible visant à assister le chercheur dans son travail d'analyse thématique des textes. Le défi principal de ce projet réside donc dans l'opérationnalisation de l'analyse thématique en employant certaines stratégies de classification et de catégorisation automatiques des textes.

Au niveau cognitif, nous proposons d'explorer la pertinence et la fécondité de certaines théories d'inspiration linguistique et littéraire ayant abordé la question du thème pour nous aider dans l'identification du contenu thématique et l'analyse thématique des documents textuels. À ce niveau, notre objectif est de démontrer comment les théories retenues, celles de Kintsch et Van Dijk, de Rimmon-Kenan et de Rastier, ont défini le thème de telle sorte qu'il est possible d'en assister informatiquement l'identification et l'analyse à l'aide de la méthodologie que nous proposons.

Au niveau informatique, un premier volet de notre démarche consiste à explorer et à comparer les performances des opérations de catégorisation et de classification automatiques à des fins d'identification du contenu thématique et d'analyse thématique des documents textuels non structurés. Les résultats sont évalués en appliquant un système de catégorisation hybride neuro-flou et un algorithme de classification neuronal non supervisé sur un corpus d'articles de journaux.

Par ailleurs, la classification et la catégorisation sont des opérations traditionnellement appliquées à des documents entiers. Nous proposons une manière alternative de réaliser ces processus : notre démarche consiste d'abord à segmenter chacun des documents puis à soumettre aux processus de regroupement les différents segments de texte. Cette démarche a l'avantage de pouvoir attribuer plusieurs catégories thématiques à chaque document, ce qui est plus difficilement réalisable lorsque les documents sont traités en entier.

Finalement, dans bon nombre d'applications d'analyse et de gestion des documents textuels, le processus de catégorisation est effectué en utilisant un plan de classification ou une taxinomie de catégories prédéfinies. Le développement de ces taxinomies, bien qu'il puisse être assisté dans certains cas par des applications informatiques, s'avère coûteux et très complexe. Dans ce projet, nous démontrerons qu'il est possible, en l'absence de taxinomies, d'employer certains termes du lexique initial du corpus comme étiquettes thématiques.

Mots-clés : analyse thématique, identification de thèmes, Lecture et Analyse de Textes Assistées par Ordinateur (LATAO), classification automatique, catégorisation automatique.

INTRODUCTION

Depuis les dix dernières années, on observe une hausse considérable du nombre d'initiatives visant à numériser et à rendre disponible, souvent à partir de l'Internet ou d'un intranet, le patrimoine informationnel des organisations et des différentes branches du savoir. Les conséquences découlant de ces initiatives sont importantes et très nombreuses : développement de normes d'encodage (XML, TEI, Dublin Core, etc.), de moyens de diffusion de l'information (portails, systèmes de gestion de contenu, etc.). Parmi ces conséquences, on note surtout une transformation radicale dans les pratiques des individus devant manipuler l'information, d'une manière ou d'une autre. Ce changement est particulièrement observable dans les pratiques des spécialistes de l'analyse et de la gestion de l'information.

On remarque, en outre, que les conséquences de ces initiatives de numérisation ont aussi des répercussions directes sur le développement des applications visant à assister l'analyse et la gestion des connaissances. Ainsi, ces initiatives visant à numériser l'information, desquelles découlent de nouvelles pratiques de gestion des connaissances et de nouvelles applications informatiques permettant d'assister les spécialistes de la gestion de l'information numérique, ont donné lieu à un nouveau territoire de recherche multidisciplinaire portant le nom de « gestion (informatisée) des connaissances » (*knowledge management*). Ce territoire peut être défini de plusieurs manières; cependant, la définition proposée par Weigel (2004), qui se situe dans une perspective informatique, nous semble des plus pertinentes :

> *Knowledge management is concerned with organizing knowledge repositories (databases, etc.) so as to allow for easy retrieval and exchange of the information stored therein. Important concepts in knowledge management include domains, i.e. fields of related concepts and terms, and ontologies, i.e. structures (typically hierarchies or networks) of interrelated terms for things, concepts, relationships, etc. in a given domain.*

Il découle de cette perspective que les applications informatiques liées à l'analyse et à la gestion informatisée des connaissances doivent être sensibles, de manière générale, à trois

dimensions précises : 1) la génération (ou la découverte), 2) la codification et 3) le transfert des connaissances. Comme le souligne Ruggles (1997, p. 3; cité dans Despres et Chauvel, 2000, p. 115), « *knowledge management tools are technologies, broadly defined, which enhance and enable knowledge generation, codification, and transfer.* »

Plusieurs spécialistes de la gestion des connaissances soulignent, à juste titre, l'importance que l'on doit de plus en plus accorder à l'analyse et à la gestion informatisées des documents textuels dans le cadre de la gestion des connaissances. Ce volet important de la gestion des connaissances intègre, par ailleurs, des concepts et des techniques provenant de plusieurs disciplines parmi lesquelles on retrouve le repérage de l'information (*information retrieval*), l'intelligence artificielle et l'apprentissage machine (*machine learning*), la linguistique computationnelle et le Traitement Automatique du Langage (TAL), les sciences de l'information et les sciences cognitives. Les manifestations concrètes des recherches sur l'analyse et la gestion des données textuelles se présentent, d'ailleurs, sous plusieurs formes. Ainsi, durant les dernières années, ont été développés des prototypes et des applications informatiques robustes visant entre autres à assister la rédaction de résumés de documents textuels (Hovy et Radev, 2002; Torres-Moreno *et al.* 2002; Mani, 2001; Mani et Maybury, 1999; COPERNIC SUMMARIZER (www.copernic.com); OPEN TEXT SUMMARIZER (www.opentext.com)), la recherche et le repérage de l'information (Jackson et Moulinier, 2002; Baeza-Yates et Ribeiro-Neto, 1999; COPERNIC DESKTOP SEARCH (www.copernic.com); DTSEARCH (www.dtsearch.com); GOOGLE DESKTOP SEARCH (www.google.com); SMART, etc.), le routage et le filtrage de l'information textuelle (Jackson et Moulinier, 2002; Hule *et al.* 1996; AUTONOMY IDOL SERVER (www.autonomy.com); IBM DB2 CONTENT MANAGER (www-306.ibm.com/software/data/cm/)), la création et la mise à jour des taxinomies (Zhang et Lee, 2004; Krishnapuram et Kummamuru, 2003; Spangler et Kreulen, 2002; INXIGHT SMARTDISCOVERY (www.inxight.com); VERITY COLLABORATIVE CLASSIFIER (www.verity.com)) et, plus récemment, des ontologies (Dittenbach *et al.* 2004; Shamsfard et Barforoush, 2004; Staab et Studer, 2004; IODE (www.ontologyworks.com), PROTÉGÉ (protege.stanford.edu), etc.).

Malgré la diversité des applications d'analyse et de gestion informatisées des documents textuels, on constate qu'un axe de recherche partagé par l'ensemble des disciplines reliées à cette problématique réside dans la compréhension et dans l'informatisation des processus

d'identification des contenus thématiques et d'analyse thématique (Louwerse et Van Peer, 2002). Comme le souligne Popping (2000, p. ix), « [thematic analysis] *is the kind of analysis that is still applied most today.* ». Actuellement, plusieurs outils visent à gérer les documents sans cependant accéder à leur contenu. Afin de dépasser les limites de ces outils, il importe de développer des techniques permettant d'accéder et de tenir compte du contenu des documents. L'identification des contenus thématiques et l'analyse thématique sont deux opérations cognitivement complexes qui ne peuvent être réalisées sans directement faire appel au contenu sémantique et informatif des textes. Nombreux sont les chercheurs qui, à l'instar de Popping, rappellent l'importance de l'analyse thématique des documents. Cependant, comme le soulignent Davi *et al.* (2005, p. 89), l'informatisation des processus d'identification des thèmes et l'analyse thématique demeurent une problématique de recherche des plus complexes (« *Extracting themes from a dataset is the most challenging task in analyzing qualitative data* »).

Il est par ailleurs intéressant de constater que la majorité des applications d'analyse et de gestion des documents intègrent des modules visant à classifier et à catégoriser automatiquement les données textuelles. Comme plusieurs auteurs l'ont souligné (Meunier, Forest et Biskri, 2005; Jackson et Moulinier, 2002; Sebastiani 2002), les processus de classification et de catégorisation automatiques semblent être au cœur de plusieurs applications d'assistance à la gestion des connaissances. Comme en témoigne la littérature scientifique et technique à ce sujet, plusieurs méthodes informatiques permettent d'accomplir automatiquement et efficacement de tels processus. Parmi les plus fréquemment citées figurent les méthodes basées sur les réseaux de neurones artificiels, la méthode des k-moyens (*k-means*), les algorithmes d'induction de règles, etc. Actuellement, un important effort de recherche est consacré à l'exploration de techniques hybrides s'inspirant des principes de la logique floue (Nauck, 1999; Ruiz et Srinivasan, 1998).

Notre projet doctoral aborde la problématique de l'identification des thèmes et de l'assistance à l'analyse thématique des documents textuels. Le projet vise à effectuer un arrimage entre des concepts et des méthodologies provenant, d'une part, des recherches linguistiques et cognitives sur l'analyse thématique et, d'autre part, des recherches appliquées en informatique sur la classification et la catégorisation des données (intelligence artificielle, apprentissage machine, etc.) afin de valider une méthodologie et un prototype d'application

permettant d'assister le chercheur spécialiste dans l'analyse et la gestion des documents dans son travail d'analyse thématique des textes.

Cette thèse est divisée en quatre principaux volets. Dans un premier temps (chapitre 1), nous présenterons brièvement la problématique dans laquelle s'insère notre projet de recherche. À cet égard, nous justifierons la pertinence et la nécessité des technologies visant à assister l'identification des thèmes et l'analyse thématique. De plus, nous identifierons les particularités respectives de chacune de ces tâches. Par ailleurs, nous présenterons les principales réalisations dans ces domaines et identifierons les limites des technologies existantes. Pour conclure ce chapitre, nous définirons les objectifs spécifiques de notre projet de recherche.

Une partie très importante de notre thèse (chapitre 2) est consacrée à la présentation de la démarche méthodologique de notre projet. Nous exposerons d'abord quelques considérations théoriques sur l'analyse thématique. Nous expliquerons pourquoi une application informatique d'assistance à l'analyse thématique se doit d'intégrer certains travaux issus des domaines de la psycholinguistique, de la sémantique et de la linguistique textuelle. À cet égard, nous présenterons le cadre conceptuel, ainsi que les différents travaux théoriques portant sur les problématiques du thème et de l'analyse thématique sur lesquels repose notre démarche. Au niveau informatique, nous présenterons les modèles théoriques employés dans ce projet ainsi que l'architecture informatique à laquelle nous avons eu recours. Nous verrons comment l'assistance à l'identification des thèmes et à l'analyse thématique peut être réalisée en employant certaines techniques provenant des domaines de l'intelligence artificielle et de l'apprentissage machine. Comme nous le verrons, les deux principales techniques employées dans ce projet sont la classification et la catégorisation automatiques dans leur application au traitement des documents textuels.

La troisième partie de ce document (chapitre 3) sera consacrée à la présentation de notre corpus d'expérimentation et des différents résultats obtenus.

La quatrième partie (chapitre 4) consistera à démontrer comment, sur la base des résultats empiriques présentés dans le chapitre précédent, il est possible d'assister informatiquement les opérations d'identification des thèmes et d'analyse thématique d'un corpus d'articles de journaux.

Finalement, nous conclurons notre travail en identifiant les avantages et les limites de notre démarche. Nous évoquerons aussi quelques perspectives d'application de nos travaux, ainsi que

les différents problèmes laissés en suspens dans ce travail et sur lesquels nous souhaitons nous attarder lors de nos recherches à venir.

CHAPITRE 1

PROBLÉMATIQUE ET OBJECTIFS DE RECHERCHE

> L'objectif poursuivi consiste à appliquer aux textes retenus un traitement permettant d'accéder à une signification non immédiatement visible (notamment par le biais de dénombrements) qui – tout en le présentant sous une forme différente – n'en dénature pas le contenu initial, mais réponde également aux questions de la problématique. D'où une double exigence de fidélité et d'originalité.
>
> Robert et Bouillaguet, 1997, p. 27.

Les domaines de l'Analyse et de la Gestion de l'Information Textuelle (AGIT) et, de manière plus ciblée, de la Lecture et de l'Analyse de Textes Assistées par Ordinateur (LATAO) sont de nature pluridisciplinaire. Ces domaines ont donc été le lieu de travaux théoriques et de réalisations informatiques des plus variés.

1.1. Caractéristiques et limites des principaux travaux d'analyse et de gestion de l'information textuelle

Dans le domaine de la recherche et du repérage de l'information[1], lequel possède déjà un

---

[1] Comme le soulignent Jackson et Moulinier (2002, p. 64), le repérage de l'information est composé de deux processus, à savoir 1) l'indexation et 2) la recherche des documents (repérage = indexation + recherche). La recherche d'information est, quant à elle, composée de trois processus. En effet, elle consiste dans 1) le traitement d'une requête, 2) l'appariement de la requête avec certains documents et 3) le tri (souvent selon la pertinence) des résultats obtenus (recherche = traitement de la requête + appariement + tri).

glorieux passé[2], les recherches théoriques et les applications informatiques sont des plus nombreuses. Parmi les travaux les plus fréquemment cités dans ce domaine figurent évidemment ceux de Salton et McGill (1983) et de Salton (1989). Ces derniers ont entre autres donné lieu au développement du système SMART fondé sur le modèle vectoriel dans son application au traitement des documents[3]. Encore aujourd'hui, ce domaine constitue un important territoire de recherche. Les recherches actuelles sont davantage consacrées à raffiner les modèles proposés durant les dernières années. Par conséquent, peu de prototypes académiques novateurs ont été proposés récemment.

Au niveau des applications informatiques robustes destinées à la recherche et au repérage de l'information, les réalisations sont de plus en plus nombreuses. Elles constituent même une partie importante du marché des applications de gestion électronique des documents (GÉD). À cet égard, les applications commerciales les plus employées sont CONVERA RETRIEVALWARE (www.convera.com), COPERNIC DESKTOP SEARCH (www.copernic.com), DTSEARCH (www.dtsearch.com), FAST DATA SEARCH (www.fastsearch.com), GOOGLE DESKTOP SEARCH (www.google.com), INDEXENGINES (www.indexengines.com), VERITY ULTRASEEK (www.verity.com) et VIVISSIMO VELOCITY (www.vivissimo.com).

Par ailleurs, le domaine de l'AGIT ne se limite pas exclusivement aux applications de recherche et de repérage de l'information. En effet, on retrouve de plus en plus d'initiatives recherches et d'applications visant à assister différentes tâches d'analyse et de gestion de l'information textuelle. En effet, de plus en plus d'efforts sont consacrés au développement d'applications commerciales visant à assister l'encodage et la validation des documents (XMLSPY (www.altova.com), XMETAL (www.softquad.com), etc.), au traitement linguistique des données textuelles (NSTEIN TECHNOLOGIES (www.nstein.com), TEMIS (www.temis.com)), à l'extraction et à la découverte d'informations (CLEARFOREST ANALYTICS (www.clearforest.com), INSIGHTFUL MINER (www.insightful.com), POLYANALYST (www.megaputer.com), etc.), à la traduction automatique (SYSTRAN (www.systransoft.com), TRADOS (www.trados.com), etc.). Au niveau de la recherche

---

[2] Bien que les principales réalisations dans le domaine de la recherche et du repérage de l'information furent réalisées depuis les années quatre-vingt, les premiers travaux fondateurs dans ce domaine, datant des années soixante, sont ceux de Luhn (1957) et de Maron (1961).

[3] Pour une excellente présentation du modèle vectoriel dans son application au traitement des documents, nous référons les lecteurs à l'article de Memmi (2000).

académique, les efforts semblent conjugués au développement d'applications pour la création et la gestion des ontologies (Staab et Studer, 2004; Missikoff *et al.* 2003; Ding et Foo, 2002a, 2002b) et des taxinomies (Zhang et Lee, 2004; Chuang et Chien, 2003; Krishnapuram et Kummamuru, 2003; Spangler et Kreulen, 2002).

Malgré la diversité et la fécondité de ces différentes applications, on constate que la majorité d'entre elles n'offrent qu'un ensemble plutôt limité de fonctionnalités permettant d'analyser véritablement le contenu des documents. Ainsi, ces logiciels sont très souvent composés de fonctionnalités principalement centrées sur des tâches de gestion de données textuelles (au détriment de l'analyse). Comme le dit Kelle (1997), « *[these software] do not provide a totally different logic of textual data management, but only more or less complicated extensions of code-and-retrieve facilities.* »

Afin de pallier à certaines limites inhérentes aux applications principalement destinées à la gestion des documents textuels, de plus en plus d'initiatives de recherche – certaines faisant même déjà l'objet de prototypes commerciaux – tentent de développer des applications permettant d'accéder au contenu informationnel des documents. En effet, au delà de certaines technologies de gestion des documents fondées presque exclusivement sur le lexique des documents (le contenu informationnel transcende l'ensemble des mots présents dans un document), un besoin pour des outils d'assistance à l'analyse du contenu thématique des documents se fait de plus en plus sentir. Plusieurs types d'applications peuvent être envisagés d'afin d'accéder au contenu des documents (dont la thématique n'est qu'une dimension précise). L'assistance à l'identification des thèmes (*topic spotting*) et à l'analyse thématique des documents semblent, à cet égard, des plus fécondes. En effet, comme l'ont clairement souligné Aery *et al.* (2003, p. 4), « *with the amount of textual data that is available and exponentially increasing there is a need to automatically process the same. One way of doing this is by topic identification, which is the process of assigning one or more labels to text.* »

1.2. L'identification des thèmes

L'identification automatique des thèmes est une tâche classique dans le domaine de l'analyse des données textuelles. Elle consiste à identifier, comme son nom l'indique, le contenu thématique (i.e. les sujets ou les thèmes) des documents. Dans une perspective

informatique il s'agit, traditionnellement, d'attribuer aux documents une ou plusieurs étiquettes ou catégories thématique (souvent à partir d'un ensemble prédéfini de catégories). Au niveau informatique, les premiers travaux dans ce domaine remontent aux années 60 (avec les travaux de Maron (1961)), mais ce n'est qu'au début des années 90 que l'assistance à l'identification des thèmes a véritablement pris son essor. Le processus d'assistance à l'identification des thèmes est une tâche informatique de nature prédictive. Dans cette perspective, elle consiste à prédire la ou les catégories thématiques dans lesquelles chaque document doit être catégorisé. L'identification thématique est donc très étroitement associée à la catégorisation automatique. Comme l'indique Sebastiani (2002, p. 1), « *text categorization (a.k.a. text classification, or topic spotting), the activity of labelling natural language texts with thematic categories from a predefined set, is one such task* ». D'ailleurs, nombreux sont les auteurs pour lesquels l'identification thématique est synonyme de catégorisation automatique[4], comme en témoigne l'extrait suivant : « *The goal in text categorization is to classify the topic or theme of a document.* » (Manning et Schütze, 2000, p. 575).

L'identification des thèmes est une tâche qui a été largement étudiée dans plusieurs projets de recherche. La littérature à ce sujet est donc vaste et abondante. Les principaux travaux dont nous nous inspirons dans notre projet de recherche sont les suivants : Sebastiani (2005b, 2005a, 2002, 1999; présentation exhaustive de l'état de l'art sur la catégorisation automatique à l'aide de techniques d'apprentissage machine), Cardoso-Cachopo et Oliveira (2004; travaux sur la comparaison de différentes méthodes (SVM et LSA) pour la catégorisation automatique des documents), Kim et Kim (2004; travaux sur la segmentation des documents pour l'identification des thèmes), De Pasquale et Meunier (2003; travaux sur l'utilisation du perceptron multicouches pour la catégorisation), Dejun et Maosong (2003; travaux sur la pondération des termes et la réduction de dimensionnalités), Han-Joon et Lee (2003; travaux sur l'utilisation de relations floues pour la construction de hiérarchies de thèmes), Haruechaiyasak *et al.* (2002; travaux sur l'utilisation d'associations floues pour la classification et la catégorisation), Joachims (2002; travaux sur l'application des SVM à des fins de catégorisation des documents), Lee *et al.* (2002; travaux sur la comparaison de

---

[4] Cependant, il nous semble important de distinguer les processus d'identification thématique et de catégorisation automatique. En effet, le premier de ces processus peut être réalisé de plusieurs autres manières alors que le second peut être appliqué à plusieurs fins autres que l'identification thématique.

différentes techniques (kNN et Rocchio) de catégorisation automatique), He *et al.* (2000; travaux sur la comparaison de différentes méthodes (kNN, SVM, ARAM) pour la catégorisation automatique des documents), Jain *et al.* (2000; présentation exhaustive de l'état de l'art sur la classification des données), Nauck (1999; travaux sur l'utilisation d'un réseau hybride neuro-flou pour la catégorisation), Yang (1999; travaux sur la comparaison de différentes techniques (kNN, LLSF, WORD) de catégorisation automatique), Yang et Lui (1999; travaux sur la comparaison de différentes techniques (SVM, kNN, ANN, LLSF, NB) de catégorisation automatique), Ruiz et Srinivasan (1998; travaux sur la comparaison de deux réseaux neuronaux (réseaux à contre-propagation et à rétropropagation) pour la catégorisation automatique), Yang et Pederson (1997; travaux sur la comparaison de techniques (DF, IG, MI, CHI, TS) de réduction de dimensionnalité), Lewis et Ringuette (1994; travaux sur la comparaison de différentes techniques (classifieur bayésien et arbres de décision) de catégorisation automatique), Lewis (1991; travaux sur la comparaison des différentes méthodes d'évaluation du processus de catégorisation).

1.3. L'analyse thématique

Bien que la problématique de l'analyse thématique puise ses racines dans les travaux de Platon et d'Aristote (Sollors, 1993), plusieurs des principales recherches concernant cette problématique furent réalisées au cours du 20^{e} siècle. Parmi les principales réalisations contemporaines qui ont fortement modélisé et contribué à la compréhension et à l'analyse des thèmes figurent celles de Tomashevsky (1925) sur la consistance et la décomposition (causale et temporaire) des thèmes en motifs (ces derniers constituant les plus petits composants d'un thème), de Aarne et Thompson (1928) sur les index de motifs, de Thompson (1946) sur la classification des motifs, de Propp (1928) sur la nécessité de développer une méthode formelle et structurée de segmentation basée sur une approche multidisciplinaire, afin d'identifier les différents motifs présents dans un texte, etc. Ces travaux fondateurs constituent les assises théoriques sur lesquelles reposent les récentes contributions à l'analyse thématique, lesquelles ont aussi largement été influencées par la distinction élaborée par l'école linguistique de Prague (principalement grâce aux travaux de N. S. Troubetskoï, de R. Jakobson et de S. O. Kartsevski) entre les concepts de « thème »

(l'objet du discours) et de « rhème » (l'information relative au thème).

Ce qui caractérise toutefois les principaux travaux actuels portant sur la question de l'analyse thématique concerne le niveau d'analyse, la manière dont les auteurs ont fréquemment abordé la question. En effet, comme le soulignent entre autres Louwerse et Van Peer (2002), Rastier *et al.* (1995) et Prince (1985), tant en littérature qu'en linguistique, la majorité des travaux ont traditionnellement abordé la problématique du thème à partir d'analyses phrastiques en tentant d'identifier les éléments, principalement linguistiques, permettant de comprendre l'organisation thématique des textes.

Malgré la richesse de la perspective linguistique, les travaux de Van Dijk et Kintsch (1983), Kintsch et Van Dijk (1978) et Van Dijk (1972) ont aussi mis en relief une approche fondée sur la distinction entre les niveaux microstructurel (reposant, de manière générale, sur l'analyse de la phrase) et macrostructurel (dont l'objet d'analyse réside dans l'ensemble du texte, pris comme un tout cohérent et structuré) du texte. Cette approche, fondée sur des travaux provenant de plusieurs disciplines des sciences humaines, semble des plus fécondes. Selon cette perspective, l'analyse thématique des textes relève d'un effort d'abstraction se situant au niveau macrostructurel et est régie par trois principes : 1) la suppression des propositions non-pertinentes, 2) la généralisation des propositions retenues et 3) l'intégration des propositions dans un tout structuré et cohérent (Kintsch, 2002 ; Louwerse et Van Peer, 2002 ; Van Dijk et Kintsch, 1983). Cette théorie permet de contextualiser l'identification et l'interprétation des thèmes d'un texte et a, en outre, motivé l'application de la théorie de la sémantique latente, des calculs de cooccurrences et de corrélation au domaine de l'analyse thématique afin d'assister le chercheur dans l'interprétation des textes auxquels il est confronté.

On constate également qu'à travers leurs diverses perspectives d'analyse, les récents travaux dans le domaine de l'analyse thématique ont clairement démontré l'importance de la variété de points de vue que peuvent apporter les différentes disciplines concernées par cette problématique. Comme le soulignent Louwerse et Van Peer (2002, p. 9), « *it seems more likely that in many disciplines thematics has always occupied a place but we may not have recognized it as such.* » Suivant Rimmon-Kenan (1985), il semble donc nécessaire d'opter pour une approche pluridisciplinaire fondée non pas exclusivement sur des considérations linguistiques, mais plutôt sur une perspective plus large permettant de tenir compte tant des phénomènes linguistiques en jeu dans le texte que des phénomènes relevant de la textualité

(processus discursif, pragmatique, etc.) sur lesquels reposent l'organisation et la structure des divers thèmes d'un corpus. Comme le soutient Giora (1985) : « la notion de thème (*topic*) discursif est indépendante de la notion de thème (*topic*) phrastique ».

Ceci témoigne bien de la distinction réelle entre, d'une part, l'analyse thématique et, d'autre part, l'identification des thèmes. Bien qu'elle soit en partie composée de ce dernier processus, l'analyse thématique ne saurait s'y réduire. Comme l'a souligné Chaar (2003, p.3), « l'identification thématique est la partie de l'analyse thématique visant à déterminer le thème d'une unité textuelle » (Chaar, 2003, p. 3).

Les recherches fondamentales sur l'analyse thématique des données textuelles ont mené à l'élaboration de plusieurs conceptions de ce type d'analyse bien spécifique. La majorité des conceptions proposées se distinguent, du moins en partie, sur la manière dont elles définissent le concept de « thème ». En effet, comme le note Rastier, plusieurs définitions de ce concept ont été proposées (« Selon qu'on privilégie le signe ou le texte, et dans le signe, le signifiant ou le signifié, le thème peut se définir par diverses voies » (2001, p. 196)). Ainsi, pour Martin (1995, p.14), le thème d'un texte est « ce à propos de quoi le texte a été composé [*aboutness*]». Cette manière d'aborder le thème est certes réductrice, mais elle demeure néanmoins partagée par plusieurs chercheurs (Rimmon-Kenan, 1985). En contre partie, il est possible de concevoir le thème selon plusieurs autres perspectives (linguistique, intentionnelle, informationnelle, etc.) (voir Rimmon-Kenan, 1985). Pour Rastier, par exemple, le thème, défini dans une perspective sémantique – laquelle « relève de la linguistique du texte et ne confère pas de prééminence à un mot-vedette identifié par son signifiant » – est spécifié « au sein de réseaux de récurrences et de transformations. »

La spécificité qui semble caractériser le processus d'analyse thématique (par opposition à l'identification des thèmes d'un corpus) réside principalement dans l'identification de la structure et des liens possibles entre les différents thèmes.

> [On peut concevoir] une quelconque manifestation du thème comme la manifestation d'un autre, et de parcourir ainsi, de proche en proche, toute la série des variations du thème [et, nous faisons l'hypothèse, des multiples thèmes qui lui sont directement ou indirectement reliés]. (Bremond, 1985, p. 419)

> À l'examen des terrains de chaque thème s'ajoute celui de leur(s) interaction(s), et les uns et les autres diffèrent selon les places et rôles respectifs des thèmes considérés. (Martin, 1995, p. 22)

C'est précisément dans cette tâche caractéristique du processus d'analyse que réside l'essentielle de sa complexité. Martin (1995, p. 17) est d'ailleurs des plus claires à ce sujet : « [...], il est relativement facile de tenir et d'entretenir un inventaire des thèmes traités d'une façon ou de l'autre. II est beaucoup plus difficile de saisir les relations que peuvent avoir les thèmes entre eux. » Cette difficulté peut être justifiée par plusieurs motifs. Cependant, l'un des facteurs principaux pouvant expliquer la complexité réside dans la subjectivité inhérente au processus d'analyse thématique.

> À qui entame l'étude d'un thème, il serait utile de connaître d'emblée le réseau notionnel dans lequel peut entrer ce thème, ce qui concerne le thème et ce qui ne le concerne pas. La question n'a pas de réponse satisfaisante et ne peut en avoir pour plusieurs raisons. L'une d'elles est qu'il ne peut y avoir consensus sur les relations qui peuvent unir un thème à un autre : le thème de l'avortement *volontaire* pourra être relié au thème du *crime* pour les uns, au thème de la *libération,* probablement *spécifique* de celui de la *femme,* pour les autres. Le thème de l'excentricité participe du thème de la *folie* pour les uns, de *l'originalité* pour les autres. Il est vrai d'ailleurs que nulle convention ne fournit de critères fermes pour déterminer le voisinage sémantique de deux thèmes. (Martin, 1995, p. 17)

Cette thèse est d'ailleurs très fréquemment soutenue dans la littérature sur l'analyse thématique. Ainsi, Martin (1995, p. 23) souligne à nouveau cette complexité en affirmant que :

> [...] la démarche thématique est multiple et complexe. Elle est avant toute chose la démarche triviale d'encodage de tout producteur de texte qui, de propos délibéré, ou sous l'effet d'une stimulation extérieure délimite le champ de sa réflexion et de son discours, à moins qu'il soit seulement guidé par une inspiration plus ou moins identifiée. Le lecteur-décodeur, et à plus forte raison le critique, tente, consciemment ou non, de distinguer les thèmes traités – annoncés ou cachés –, s'approprie les concepts-clés le temps du commentaire, les transfère nécessairement dans son univers, même s'il prétend rester dans le cadre étroit d'un écrivain, d'un temps, d'un lieu donnés, et les livre à son propre lecteur qui, à son tour... C'est probablement dans cette flexibilité qui est en même temps richesse que réside la véritable originalité de la notion de « thème ».

Bref, l'analyse thématique doit donc être perçue comme un processus manifestement complexe étroitement lié à l'activité interprétative de parcours et de lecture des textes.

Contrairement à l'identification thématique, peu d'applications informatiques destinées exclusivement à l'assistance à l'analyse thématique ont été proposées. Les seules véritables applications – issues de projets de recherche académiques, pour la plupart, sont celles de

Roy et Beust (2004; logiciel PROXIDOCS permettant d'identifier les différentes propriétés thématiques d'un corpus), Havre *et al.* (2002; logiciel THEMERIVER pour la visualisation des variations thématiques), de Hogenraad (2002; logiciel PROTAN pour la description des thèmes d'un corpus), de Kastberg Sjöblom et Brunet (2000; logiciel HYPERBASE pour la description des thèmes d'un corpus) et de Miller *et al.* (1998; logiciel TOPIC ISLANDS pour la visualisation et l'exploration des thèmes). Finalement, il existe aussi quelques logiciels destinés à l'analyse de contenu pouvant être partiellement utiles afin d'identifier et de décrire les contenus thématiques de corpus.

Les travaux les plus comparables aux nôtres sont manifestement ceux de Rossignol (2005) et de Rossignol et Sébillot (2005, 2003, et 2002). Ces travaux, menés presque simultanément aux nôtres, abordent principalement le problème de l'extraction de mots-clefs thématiques. L'objectif de ces travaux consiste à identifier la présence de thèmes dans des segments de textes sur la base de cooccurrences de mots-clés. De manière générale, la démarche privilégiée par ces auteurs consiste à employer d'abord un algorithme de classification hiérarchique sur des segments de documents, puis à calculer la cooccurrence entre les mots de chaque regroupement pour ne retenir, en dernier lieu, que les mots dont la cooccurrence est la plus élevée. Comme nous le verrons, l'approche proposée par Rossignol et Sébillot est partiellement comparable à certains volets de la démarche que nous proposons dans ce document. Cela témoigne de toute évidence de la pertinence d'une approche fondée sur la classification et l'extraction automatique de termes permettant d'identifier le contenu thématique de documents textuels.

Nos travaux se distinguent cependant de ceux de Rossignol et Sébillot à plusieurs égards. D'une part, les deux démarches que nous proposons reposent sur des techniques de regroupement (le réseau neuronal ART1 et le réseau hybride neuro-flou intégré dans l'application NEFCLASS-J) dont très peu de chercheurs ont exploré la pertinence dans le cadre de traitements automatiques de documents textuels. D'autre part, notre démarche est réalisée en segmentant les documents de notre corpus initial. L'opération de segmentation est réalisée afin d'atteindre un objectif bien spécifique. En effet, cette opération permet d'identifier facilement la variété de thèmes que l'on peut retrouver dans un seul document. Il s'agit donc d'une opération permettant d'optimiser plusieurs opérations liées à la gestion électronique des documents. Par ailleurs, une partie importante de notre travail consiste à identifier

explicitement les fondements théoriques (concernant principalement le concept de « thème ») sur lesquels reposent l'ensemble de notre démarche informatique. Finalement, contrairement aux travaux de Rossignol et de Sébillot, l'objectif ultime de notre travail consiste à valider une démarche méthodologique permettant d'assister l'analyse thématique des documents textuels. Dans cette démarche, l'opération visant à identifier le contenu thématique des segments de documents ne constitue qu'une étape particulière. Cette étape est très importante, mais elle s'insère dans un processus beaucoup plus complexe d'assistance à la découverte, au parcours et à l'analyse thématique des documents textuels.

1.4. Limites de la technologie

Dans les domaines de la LATAO, du forage des textes et de la GÉD, le problème central de la technologie existante pour l'identification des thèmes et l'analyse thématique est qu'elle ne permet que très partiellement l'identification automatique des thèmes ou des catégories thématiques et donc encore moins la découverte des structures et des réseaux thématiques qui leur sont associés. En effet, la majorité des logiciels, très souvent inspirés des travaux dans le domaine de l'analyse de contenu (*content analysis*) (WORDSTAT, GENERAL INQUIRER, etc.), permettent d'effectuer plusieurs calculs statistiques complexes sur les catégories thématiques, mais reposent sur un processus de catégorisation thématique où chaque terme d'un corpus doit être manuellement associé à une catégorie thématique.

> *The researcher, however, needs first to make sense of the set of common terms in each group or cluster of responses. The researcher then needs to read all the responses and decides on the fit of each response under a specific theme. Hence, while the software helps in locating and counting words and clustering responses, the researcher needs to do most, if not all, of the time-consuming work.* (Davi *et al.*, 2005, p. 102).

> L'analyse lexicale, dont la statistique est un auxiliaire, ne propose pas d'elle-même des indices à l'analyse thématique. [...] En d'autres termes, pour atteindre ses objectifs, la thématique doit guider l'analyse lexicale, puis interpréter ses résultats qui sans cela resteraient inutilisables pour une sémantique textuelle. Les logiciels d'interrogation imposent certaines démarches, mais ne proposent rien. Ils servent à confirmer ou infirmer des hypothèses, qui dépendent de la stratégie d'interprétation. (Rastier, 1995, p. 223)

Autrement dit, sauf pour quelques outils qui assistent l'attribution manuelle ainsi que la manipulation des catégories (WORDSTAT, NUD*IST, ATLAS/TI, etc.), les technologies existantes, comme le soulignent plusieurs critiques (Barry, 1998; Kelle, 1997), n'offrent que peu d'outils pour faire émerger les catégories thématiques elles-mêmes, leur étiquetage catégoriel et, surtout, les différents réseaux thématiques présents dans les documents. Ainsi, les logiciels s'arrêtent souvent là où le travail d'analyse thématique doit justement commencer. Or dans ce type d'analyse ce sont les thèmes et les différentes structures thématiques qui sont l'objet de la recherche.

> *These are text analysis tasks and methodological approaches, which are not supported enough, or there is functionality needed, which has yet to be developed or integrated in text analysis software.* (Alexa et Zuell, 1999a, p. 133)

Il semble donc nécessaire de développer des outils plus sophistiqués visant à soutenir cette démarche de classification et de catégorisation thématique sur laquelle repose l'analyse thématique des documents textuels (Meunier, Forest et Biskri, 2005).

1.5. Objectifs du projet

L'objectif général de ce projet consiste à développer, à valider et à comparer deux méthodologies informatiques fondées sur la classification et la catégorisation automatiques permettant d'assister efficacement l'identification des thèmes et, surtout, l'analyse thématique de documents textuels. Il vise ainsi à effectuer un transfert de concepts et de méthodologies provenant, d'une part, des recherches théoriques et pluridisciplinaires sur l'analyse thématique et, d'autre part, des recherches appliquées en classification et en catégorisation automatiques des données afin de développer une méthodologie et un prototype d'application flexible visant à assister le chercheur dans son travail d'analyse thématique des textes. Le défi principal de ce projet réside donc dans l'opérationnalisation de l'analyse thématique en employant certaines stratégies de classification et de catégorisation automatiques des textes. L'hypothèse générale du projet est formulée ainsi : *l'identification automatique des thèmes et l'analyse thématique des données textuelles peuvent être assistées efficacement en employant certaines techniques issues de l'intelligence artificielle et de*

l'apprentissage machine à des fins de classification et de catégorisation automatiques des données textuelles.

Cet objectif général présuppose l'atteinte de trois sous-objectifs spécifiques. Comme nous y avons fait allusion précédemment, tout projet dont l'objectif est d'assister informatiquement les processus d'identification des thèmes et d'analyse thématique implique nécessairement un cadre théorique spécifique et une conception précise de ces deux processus. Cependant, plusieurs des travaux théoriques sur ces processus ne semblent pas en permettre une véritable informatisation. Par ailleurs, très peu d'études ont tenté de valider la fécondité des différentes perspectives théoriques à des fins d'informatisation de ces processus. Par conséquent, notre sous-objectif, au niveau cognitif, consiste à explorer et à mettre à contribution la pertinence de certains travaux en sémantique cognitive, en linguistique textuelle et en psycholinguistique dans leur application à l'informatisation des deux processus liés à la thématique. Plus spécifiquement, notre objectif consiste à exploiter les travaux de Van Dijk (1972) et de Kintsch et Van Dijk (1978), de Rastier (2001, 1995, 1994, 1987/1996) (mais aussi de Rastier *et al.* (1995, 1994)) et de Rimmon-Kenan (1985) à des fins d'assistance informatique à l'analyse thématique. Nous ne retiendrons de ces travaux que les aspects spécifiques pertinents à l'égard de notre problématique. Comme nous le verrons, la perspective proposée par Rastier permet de rendre compte de phénomènes thématiques qu'une approche purement linguistique ne saurait justifier de manière exhaustive. Par ailleurs, nous présenterons aussi la distinction proposée de Van Dijk et Kintsch entre les niveaux microstructurel et macrostructurel et nous verrons en quoi l'analyse thématique est perçue, dans cette optique, en tant qu'abstraction se situant au niveau macrostructurel régie par trois principes : 1) la suppression des propositions non-pertinentes, 2) la généralisation des propositions retenues et 3) l'intégration des propositions dans un tout structuré et cohérent.

Par ailleurs, comme l'ont noté Rastier (2001, 1995) et Rimmon-Kenan (1985), une approche exclusivement linguistique de l'analyse thématique implique une conception trop étroite et trop restrictive de la thématique. Ainsi, plusieurs études ont abordé le thème dans une perspective davantage discursive ou textuelle, plutôt qu'exclusivement linguistique. Ainsi, à défaut d'imposer au thème une manifestation exclusivement phrastique (relevant donc de la linguistique), cette position permet, quant à elle, de percevoir le thème comme

élément plus englobant de tout discours. À cet égard, la notion de thème discursif (Van Dijk, 1972) permet, nous semble-t-il, de concevoir le thème à une échelle plus large en tant qu'élément textuel complexe et englobant dont l'émergence se fait sur la base des sous-éléments phrastiques qui le composent (« Les thèmes peuvent être, par définition, plus abstraits ou plus généraux que les énoncés spécifiques figurant dans une phrase donnée » (Rimmon-Kenan, 1985, p. 401)). D'ailleurs, cette conception semble rejoindre celle d'importants chercheurs dans le domaine (Louwerse et Van Peer, 2002).

D'un point de vue cognitif, notre démarche visera donc à mettre en valeur la complémentarité des différentes approches abordées et à démontrer comment ces travaux permettent d'opérationnaliser informatiquement les processus d'identification du contenu thématique et d'analyse thématique à l'aide des méthodologies développées.

Au niveau informatique, notre projet est composé de deux sous-objectifs. Nous avons mentionné (voir hypothèse générale) que notre projet consiste à appliquer certaines techniques de classification et de catégorisation automatiques à des fins d'analyse thématique des documents textuels. Nous avons déjà démontré la pertinence de la classification à des fins d'analyse des données textuelles (Forest, 2002; Forest et Meunier, 2000). Dans le cadre de notre projet, la particularité de notre démarche consistera cette fois (comme nous le démontrons plus en détail dans le chapitre 2) à distinguer et à explorer la pertinence de deux techniques complémentaires de forage de textes. La première d'entre elles réside dans la catégorisation automatique des données. Il s'agit, comme nous le verrons, d'une technique de nature prédictive présupposant que nous disposions *a priori* de certaines métadonnées concernant notre corpus initial. La seconde réside dans la classification automatique des données. Contrairement à la catégorisation, la classification est une opération de nature exploratoire qui ne présuppose aucune métadonnée préalable concernant le contenu des documents à analyser.

À notre connaissance, aucun projet de recherche dans les domaines de l'AGIT et de la LATAO n'a exposé de manière explicite et exhaustive la distinction et la complémentarité entre ces deux techniques de forage de données[5] dans leur application au traitement des

---

[5] Dans le domaine du forage de données, seuls les travaux de Aggarwal *et al.* (2004) ont esquissé cette distinction et ont tenté de jumeler au sein d'un même processus les opérations de la classification et de la catégorisation à des fins de catégorisation des données textuelles. Cependant, cette démarche a été

documents textuels. Comme en témoigne la littérature dans le domaine du forage de données (*data mining*) (Weiss *et al.*, 2005, Alpaydin, 2004; Manning et Schütze, 1999), les opérations de catégorisation et de classification sont différentes et peuvent être réalisées en employant des stratégies de natures distinctes. La distinction entre ces deux opérations est aussi observable lorsque ces opérations sont appliquées au traitement des données textuelles. Dans le deuxième chapitre, nous exposerons en quoi consiste la distinction entre ces deux opérations. En outre, nous verrons aussi que ces opérations, bien que distinctes, ne sont pas incompatibles. Dans une perspective reliée au traitement des données textuelles, bien qu'elles permettent d'obtenir des résultats comparables, elles se distinguent essentiellement dans la mesure où elles ne réalisent pas des opérations cognitives identiques. L'une (la catégorisation), vise à projeter sur un ensemble de documents, via un processus d'apprentissage, des métadonnées connues ou obtenues à partir d'un autre ensemble de documents. L'autre (la classification) vise plutôt à regrouper des documents en ne faisant intervenir aucune information connue *a priori* concernant ces mêmes documents[6].

La première hypothèse secondaire, au niveau informatique, peut donc être formulée ainsi : *les processus de catégorisation et de classification automatiques, bien que de nature différentes, sont complémentaires. Leur distinction relève, d'une part, de la nature de*

---

effectuée à d'autres fins que celles que nous poursuivons dans notre projet. En effet, leur démarche consistait à appliquer d'abord le processus de classification afin d'extraire une taxinomie de catégories, puis à appliquer le processus de catégorisation afin de catégoriser les documents. Leur objectif principal était donc, *grosso modo*, de démontrer la pertinence de la classification dans la construction d'une taxinomie.

[6] Au départ, la démarche envisagée consistait à jumeler les opérations de classification et de catégorisation automatiques. Nous avons avancé l'hypothèse selon laquelle le jumelage de la classification et de la catégorisation automatiques pouvait engendrer de meilleurs résultats au niveau de la catégorisation. Dans cette perspective, ce ne sont plus des documents qui allaient être soumis au processus de catégorisation (tel que cela est généralement effectué), mais plutôt des classes de segments de documents partageant certaines caractéristiques lexicales communes. Combiné à un processus de filtrage du lexique, nous croyions que ce jumelage allait avoir comme effet d'accroître et d'uniformiser les termes de chaque élément (c'est-à-dire des classes de segments de documents) à catégoriser. Nous avons testé cette démarche. Les résultats préliminaires obtenus nous ont malheureusement menés à revoir notre démarche. En effet, bien que théoriquement possible, notre démarche initiale ne semble pas réalisable dans la pratique. En effet, l'opération de classification des données permet, entre autres, de réduire substantiellement le volume de données initial. Cependant, comme l'opération de catégorisation automatique présuppose une quantité importante de données représentatives afin d'effectuer un apprentissage, l'opération de classification s'est avéré un obstacle, réduisant ainsi dans les performances du processus d'apprentissage sur lequel repose principalement l'opération de catégorisation automatique.

l'opération qu'ils réalisent et, d'autre part, de la quantité d'informations qu'ils font intervenir concernant les données initiales. Cependant, malgré leurs différences, ces deux processus peuvent être employés pour assister informatiquement l'identification des thèmes et l'analyse thématique.

Ceci nous amène à mentionner une autre particularité de notre démarche. Traditionnellement, la classification et la catégorisation sont des opérations appliquées aux documents entiers. En nous basant sur les récents travaux de Kim et Kim (2004), nous proposons une manière alternative d'appliquer ces processus. Ainsi, notre démarche consiste d'abord à segmenter chacun des documents du corpus puis à soumettre au classifieur ou au catégoriseur les différents segments de texte. Comme l'ont démontré Kim et Kim, cette démarche possède l'avantage d'attribuer plusieurs catégories thématiques à chaque document (ce qui est plus difficilement réalisable lorsque les documents sont traités en entier).

Dans bon nombre d'applications d'AGIT et de LATAO, l'opération de catégorisation est effectuée en utilisant un plan de classification ou une taxinomie de catégories prédéfinies (Jackson et Moulinier, 2002; Manning et Schütze, 1999; Sebastiani, 1999). Cependant, comme plusieurs spécialistes l'on souligné (Boeri, 2004; Feldman, 2004), le développement des taxinomies, bien qu'il puisse être assisté dans certains cas à l'aide d'applications informatiques, s'avère très coûteux. Les taxinomies doivent être étroitement adaptées aux domaines d'application et leur construction nécessite un travail de lecture et d'analyse manuelle devant être réalisé par des spécialistes en terminologie. Dès lors, il importe de développer des solutions alternatives permettant d'assister le processus de catégorisation sans avoir recours à de telles taxinomies. Dans ce projet, nous verrons qu'il est possible, en l'absence de taxinomies, d'employer certains termes du lexique initial du corpus comme étiquettes thématiques.

La deuxième hypothèse secondaire, au niveau informatique, peut donc être formulée ainsi : *certaines mesures permettent d'extraire les termes les plus représentatifs de classes de documents. Ces termes peuvent alors être employés comme étiquette décrivant adéquatement le contenu thématique des classes auxquelles ils sont associés.*

CHAPITRE 2

MÉTHODOLOGIE

> *Documents are grouped because they are in some sense related to each other; but more basically, they are grouped because they are likely to be wanted together [...].*
>
> Van Rijsbergen, 1979, p. 24

> Il faut grouper les énonciations contenues dans chaque livre et les réduire à un certain nombre de chefs principaux, de façon à retrouver aisément toutes celles qui se rapportent au même objet.
>
> Spinoza, 1955, pp. 714-715

Nous atteindrons notre objectif principal en employant puis en comparant deux démarches méthodologiques, comportant chacune quatre étapes principales. La première comporte les opérations suivantes : 1) le prétraitement, 2) la vectorisation, 3) la catégorisation thématique et 4) la découverte et l'analyse thématique des documents. La seconde démarche est, pour sa part, constituée des opérations suivantes : 1) le prétraitement, 2) la vectorisation, 3) la classification et l'extraction des termes thématiques et 4) la découverte et l'analyse thématique des documents. Les deux démarches employées sont presque identiques : elles ne se distinguent qu'au niveau de l'opération effectuée à la troisième étape. En outre, nous comparerons, dans le quatrième chapitre, les résultats produits par ces deux démarches sur un même corpus d'expérimentation.

Comme nous l'avons mentionné précédemment, ces deux démarches méthodologiques

reposent sur des conceptions théoriques bien précises du concept de thème, que nous allons maintenant expliciter.

2.1. Thème : enjeux et cadre théoriques

Dans cette section, nous présentons les théories sur lesquelles repose notre opérationnalisation des processus d'identification du contenu thématique et d'analyse thématique. Nous procédons du plus général au plus particulier. L'objectif de cette section n'est pas de présenter de manière exhaustive l'ensemble des théories de chacun des auteurs abordés; nous voulons plutôt démontrer comment les trois points de vue exposés définissent adéquatement le concept de thème, tout en permettant une informatisation des opérations d'identification du contenu thématique et d'analyse thématique de documents textuels.

2.1.1. Thématique et macrostructure textuelle

Comme nous l'avons évoqué précédemment, la majorité des travaux ayant abordé la question de la thématique ont très souvent hérité des acquis de la tradition linguistique et ont abouti à diverses définitions du concept de thème se situant toutes à un niveau exclusivement phrastique. Cette perspective permet certes d'expliquer certains phénomènes thématiques présents au niveau de la phrase, mais elle permet difficilement de tenir compte de phénomènes thématiques attestés dont la manifestation ne saurait se réduire à un simple énoncé linguistique. Sensibles à cette distinction, Kintsch et Van Dijk (1978) ont abordé la problématique de la compréhension du discours en proposant une distinction entre les niveaux sémantiques microstructurel et macrostructurel du discours.

La microstructure se situe au niveau de la phrase; elle concerne l'organisation structurelle, ainsi que les relations entre les différentes propositions du discours (Kintsch et Van Dijk, 1978, p. 365). Dans cette optique, un texte n'est pas une suite aléatoire de propositions indépendantes. En effet, il s'agit plutôt d'un ensemble d'unités structurées et cohérentes. Selon ces auteurs, il est possible d'identifier la microstructure d'un texte en fonction de la cohérence référentielle observable entre les différentes propositions. Comme le mentionnent Kintsch et Van Dijk (1978, p. 365), « *we can establish a linear or hierarchical sequence of*

propositions in which coreferential expressions occur. The first (or superordinate) of these propositions often appear to have a specific cognitive status in such a sequence, being recalled two or three times more often than other propositions. »

Par opposition au niveau microstructurel, celui de la macrostructure du texte vise à identifier et à décrire la structure sémantique du texte ou du discours à un niveau global, plus général. C'est sur la base de cette seconde approche que repose le concept de « thème du discours » (*topic of discourse*) par opposition au thème phrastique (*sentence topic*). Kintsch et Van Dijk (1978, p. 366), soutiennent ce principe ainsi :

> « *The theoretical and linguistic reasons for this level of description derive from the fact that the propositions of a text base must be connected relative to what is intuitively called a* topic of discourse *(or topic of conversation), that is, the theme of the discourse or a fragment thereof. Relating propositions in a local manner is not sufficient. There must be a global constraint that establishes a meaningful whole characterized in terms of a discourse topic.* »

Selon ces deux auteurs, l'identification des thèmes ou du contenu thématique d'un texte relève d'une analyse se situant au niveau macrostructurel.

Dans leurs travaux initiaux, Kintsch et Van Dijk proposent d'aborder les niveaux microstructurel et macrostructurel en employant la phrase comme élément d'analyse. Cette position découle directement de l'objectif qu'ils tentent d'atteindre. En effet, ils visent à démontrer comment le thème global du texte ou du discours peut être relié aux différentes propositions composant un texte (« *In order to show how a discourse topic is related to the respective propositions of a text base, we thus need semantic mapping rules with microstructural information as input and macrostructural information as output.* » (1978, p. 366)). Le modèle qu'ils ont proposé accepte comme intrant les différentes phrases d'un texte et vise à en extraire d'abord les propositions. Ces dernières sont extraites en identifiant les concepts de chaque phrase (à chaque concept correspond une proposition). En raison d'une règle logique de compositionnalité, chaque proposition doit inclure un prédicat (un concept relationnel) et au moins un argument (pouvant être traduit en termes de fonction sémantique du type « agent », « objet », « but »). Par la suite, grâce à un processus de traitement fondé sur des schémas de transformation itératifs, une macrostructure (ou une hiérarchie de macrostructures) est engendrée. À partir de cette macrostructure, il est possible

de condenser et d'organiser le contenu des diverses unités propositionnelles. Ces processus de condensation et d'organisation sont réalisés en respectant la vérité et la signification de chacune des propositions traitées et sont régis par trois principales opérations : 1) la suppression des propositions non-pertinentes, 2) la généralisation des propositions retenues et 3) la construction de propositions plus générales respectant la vérité et la signification des propositions provenant de la microstructure généralisée.

Nous n'insistons pas davantage sur les détails théoriques du modèle proposé par Kintsch et Van Dijk, car nous croyons que cette présentation générale est suffisante afin de saisir les particularités de leur démarche que nous retenons dans le cadre de notre projet. Ainsi, des travaux de ces auteurs, nous retenons deux aspects généraux qui guideront notre méthodologie informatique.

D'une part, les travaux de Kintsch et Van Dijk attestent clairement de la distinction entre les niveaux microstructurel et macrostructurel observables dans un texte ou un corpus de données textuelles. Il existe manifestement une distinction entre le contenu thématique observable au sein d'un énoncé isolé et celui qui émerge d'une structure plus complexe composé d'un ensemble d'énoncés interreliés. Sur la base de cette distinction, nous ne nous attarderons pas au thème dans une perspective phrastique, car notre objectif consiste à cerner « ce à propos de quoi » un ensemble de segments de texte est rédigé (identification du contenu thématique) et à identifier les différents parcours possibles entre les thèmes (analyse thématique).

D'autre part, nous avons évoqué le fait que le modèle proposé intègre trois principales opérations permettant de passer du niveau microstructurel au niveau macrostructurel. Il s'agit des opérations de suppression (de l'information non pertinente), de généralisation (des propositions retenues) et d'intégration (des propositions dans un tout structuré et cohérent). Nous avons aussi vu que ces opérations étaient réalisées en employant comme intrant des propositions isolées afin de générer en sortie une macrostructure plus générale respectant la cohésion des différentes propositions. Dans notre projet, nous retenons le principe général du modèle proposé par Kintsch et Van Dijk, mais nous sommes d'avis que le processus de transformation peut être appliqué à d'autres éléments que les propositions isolées. Ainsi, comme nous le verrons dans les sections suivantes, la méthodologie que nous proposons s'inspire de ces trois principes, mais, plutôt que de les appliquer sur un ensemble de

propositions, elle les applique sur un ensemble de termes (i.e. sur le lexique) extraits de chacun des segments de texte préalablement regroupés. Nous verrons que notre démarche s'apparente à celle de Kintsch et Van Dijk, car elle vise à supprimer les différents éléments non-pertinents de notre corpus, à intégrer les différents segments de documents en les soumettant à un processus de regroupement (qu'il soit fondé sur la classification ou la catégorisation) et à généraliser les différents regroupements en leur attribuant une étiquette permettant d'en représenter adéquatement le contenu thématique.

2.1.2. Le thème dans tous ses états

Les travaux de Kintsch et Van Dijk permettent de distinguer deux conceptions du thème. La première conception, d'influence linguistique, considère le thème dans une perspective essentiellement phrastique; la seconde, inspirée de travaux dans le domaine de l'analyse du discours, l'aborde plutôt dans une perspective globale en le considérant comme caractéristique du discours. Il est cependant possible d'approfondir davantage cette distinction et de proposer différentes définitions plus nuancées du concept de « thème ». À cet égard, les travaux de Rimmon-Kenan nous apparaissent des plus pertinents.

Dans sa contribution de 1985 (traduite en anglais en 1995 dans un remarquable ouvrage édité par Bremond, Landry et Pavel), Rimmon-Kenan présente les principales caractéristiques de sa conception du thème. Ce travail, d'inspiration manifestement littéraire, s'avère des plus enrichissants dans le cadre de notre projet. L'intérêt de la contribution de Rimmon-Kenan est double : en plus de proposer une conception des plus fécondes, il dresse un inventaire concis, mais cependant fidèle, des principales conceptions de la thématique qui ont caractérisé les travaux du siècle dernier. En nous inspirant de la contribution de Rimmon-Kenan, nous présentons dans cette section un bref survol des différentes conceptions de la thématique qui ont enrichi la réflexion sur cette problématique[1].

a) Le thème linguistique. Le concept de « thème » a été introduit en linguistique par

---

[1] Pour une description et une évaluation exhaustive des différentes conceptions du thème dans une perspective phrastique, nous référons le lecteur à Reinhart (1981).

l'école de Prague[2], en tant qu'élément distinctif du « rhème ». Selon cette école, le concept de thème peut être défini comme l'objet du discours et le rhème, comme l'information relative à ce thème[3]. Il s'agit d'une importante distinction qui fut d'ailleurs reprise par Beardsley (1958) lorsqu'il oppose les notions de « thème » et de « thèse ». Cette conception linguistique du thème a donné lieu à différentes conceptions d'inspiration linguistique partageant toutes une perspective linéaire et syntaxique du thème. Rimmon-Kenan souligne à ce sujet : « *most linguistic studies deal with sentence topics [...]. A sentence topic must correspond to an expression in the sentence, and consequently linguists set themselves the task of formulating criteria for defining the topic expression.* » (p. 10). Pour certains (dont Halliday), le thème réside dans les premiers énoncés (*expression*) de la phrase. D'autres (dont les linguistes inspirés de l'École de Prague) définissent le thème comme étant le sujet de la phrase. Finalement, certains théoriciens des dernières décennies (dont Chomsky et Jackendoff) prétendent que le thème réside plutôt dans l'unité linguistique non-accentuée dans la phrase. La pluralité des points de vue témoigne assurément de l'absence de consensus sur ce qu'est un thème dans une perspective linguistique. Malgré la diversité des définitions, nous notons que toutes les hypothèses avancées partagent à tout le moins une caractéristique : le thème, dans une perspective linguistique, doit être approché en prenant comme élément d'analyse la phrase.

Comme le souligne à juste titre Rimmon-Kenan (1985, p. 399), les différentes perspectives linguistiques n'offrent malheureusement qu'une définition réduite et partielle du concept de thème.

> Un thème n'est pas une entité *dans* : ce n'est pas un segment inclus *dans* le continuum du texte, mais une construction dont l'assemblage s'effectue à partir des éléments discontinus du texte. En bref, l'unité plus large à laquelle se réfère le thème en littérature n'est pas de l'ordre de la phrase mais de celui du texte considéré comme un tout, et le texte représente plus que la somme de ses phrases. [...] Si un thème n'est pas un segment inclus dans le continuum du texte, on comprend pourquoi il ne peut être défini en termes d'ordre linéaire et de fonction syntaxique.

---

[2] L'école de Prague (ou cercle de Prague), fondé par N. S. Troubetskoï, R. Jakobson et S. O. Kartsevski en 1926, était composé de linguistes dont les théories étaient en grande partie fondées sur les travaux de Ferdinand de Saussure.

[3] Par exemple, selon cette approche, dans la phrase « Cette thèse est vraiment exceptionnelle », « Cette thèse » constitue le thème et le reste de la phrase (« est vraiment exceptionnelle »), le commentaire, le rhème.

Il semble en effet difficile d'admettre que le thème d'un document textuel (ou même d'un segment de texte) puisse se réduire à l'ensemble des thèmes linguistiques des énoncés qui le composent. Pour qu'il puisse en être ainsi, il faudrait un principe de compositionnalité sur la base duquel nous pourrions jumeler l'ensemble des thèmes linguistiques d'un corpus afin de générer un seul thème générique reflétant l'entièreté de « ce dont il est question » dans un document. Bien que certains auteurs (dont Kintsch et Van Dijk) aient abordé cette problématique des plus complexes, l'état actuel des travaux dans le domaine ne semble pas suffisamment étoffé, ce qui nous incite à remettre en question une telle perspective linguistique dans son application à l'identification informatique du contenu thématique.

b) Le thème, intention et intérêt du locuteur. Certains théoriciens (Garcia, 1975; Schachter, 1973) ont tenté d'explorer le thème en tant que reflet des intentions d'un locuteur. Dans cette perspective, le thème est le point central de l'attention du locuteur. Comme le note Rimmon-Kenan, il est curieux de constater que l'attention du locuteur a aussi été employée par d'autres auteurs afin de définir le commentaire, le rhème. En plus de mener à des interprétations contradictoires, cette perspective présuppose que nous disposions de critères permettant d'identifier clairement l'attention du locuteur. À ce jour, la définition de ces critères fait toujours l'objet de vifs débats théoriques.

c) Le thème en tant qu'information. Il est aussi possible de concevoir le thème dans une perspective informationnelle. Le thème est alors considéré comme l'« information ancienne », alors que le rhème représente l'« information nouvelle ». Nous pouvons adresser à cette définition la même critique que nous avons adressée à la conception intentionnelle du thème : la distinction entre « information ancienne » et « information nouvelle » implique la présence de critères dont la nature est toujours l'objet de débats.

d) Le thème en tant qu'« à-propos-de ». Une conception du thème en tant que « ce à propos de quoi » un document, une phrase ou un document est rédigé semble des plus intuitives. Les travaux s'inscrivant dans cette optique ont distingué l'« à-propos-de sémantique » (*semantic aboutness*) de l'« à-propos-de pragmatique » (*pragmatic aboutness*).

L'« à-propos-de sémantique » requiert aussi que nous disposions de critères, mais, plus encore, il impose, comme l'indique Rimmon-Kenan (p. 400), que « ce à propos de quoi une phrase est dite subsiste indépendamment de la diversité de ses formulations équivalentes, tandis que, selon le linguiste, des formulations différentes peuvent conduire à des thèmes différents. »

L'« à-propos-de pragmatique », proposé par Reinhart (1981), vise à proposer une solution au problème de l'« à-propos-de sémantique ». Cette conception réside dans l'élaboration de taxinomies de thèmes. D'un point de vue linguistique, cette perspective ne permet cependant pas de distinguer le thème du rhème d'un énoncé et son application à l'analyse de documents textuels ne peut être réalisée sans soulever des enjeux théoriques et pratiques importants.

e) Le thème en tant qu'étiquette. Ce survol des principales conceptions du thème a mené Rimmon-Kenan à proposer une définition générale du thème qui laisse entrevoir la possibilité d'en informatiser l'identification. Inspirée des trois opérations que nous avons retenues des travaux de Kintsch et Van Dijk, elle consiste à définir le thème en tant qu'étiquette obtenue par l'assemblage de différentes unités linguistiques observables dans les documents. Les trois opérations réalisées afin d'identifier le thème d'un document en tant qu'étiquette sont 1) l'assemblage, 2) la généralisation et 3) l'étiquetage. Rimmon-Kenan présente cette conception dans les termes suivants (p. 402) :

> Les éléments sont d'abord assemblés en un schéma élémentaire, une catégorie unificatrice de niveau inférieur établie sur la base de quelque rapport de récurrence, de similarité, de contraste ou d'implication que l'on discerne entre eux. Les étiquettes, exprimant le dénominateur commun, ressemblent au catalogue « par matières » de Reinhart, mais elles intègrent des éléments hétérogènes qui, de plus, n'ont pas besoin de fonctionner comme les thèmes en linguistique. Les catégories étiquetées sont alors reliées à d'autres catégories du même ordre d'après les mêmes principes de cohésion ; l'opération aboutit soit à l'élaboration d'une étiquette plus générale, soit à l'augmentation de la capacité d'intégration de l'étiquette initiale. Les thèmes sont les étiquettes du niveau le plus élevé trônant au sommet d'une structure hiérarchique en forme d'arbre.

En d'autres termes, un thème est une étiquette (se situant au niveau le plus élevé d'une structure hiérarchique), c'est-à-dire un dénominateur commun fondé sur un principe de cohésion (récurrence, similarité, contraste ou implication). La perspective proposée par Rimmon-Kenan implique volontairement une composante subjective importante dans le processus d'étiquetage thématique. Ce volet subjectif (dont les fondements théoriques ne sont pas explicitement évoqués par Rimmon-Kenan) est cependant inévitable. L'activité de thématisation implique fondamentalement une activité cognitive d'interprétation. Le texte offre au lecteur analyste certains indices thématiques, mais c'est au lecteur que revient, en dernière instance, l'activité d'identification du contenu thématique. Malgré cette composante subjective, il n'en demeure pas moins possible, comme nous le verrons, d'assister informatiquement cette opération.

2.1.3. Thème et sémantique textuelle

Les travaux sur la thématique auxquels nous avons fait référence dans les sections précédentes nous permettent de spécifier les caractéristiques générales du cadre théorique sur lequel repose notre méthodologie informatique (décrite à la section 2.2). Bien que fondamentaux, ces travaux demeurent cependant très généraux. Nous proposons donc dans cette section de spécifier davantage les particularités du cadre théorique employé en y intégrant certaines caractéristiques de la conception du thème développée par Rastier. Nous avons choisi d'appuyer notre démarche sur les importants travaux en linguistique et en sémantique textuelle de F. Rastier, car nous sommes d'avis que les thèses développées par cet auteur, tout en récupérant certaines des idées exposées précédemment, constituent des assises des plus solides pour l'informatisation des opérations d'identification du contenu thématique et d'analyse thématique des documents textuels.

Dans son ouvrage de 2001, Rastier aborde la question de la thématique en distinguant les différentes postures qu'il est possible d'adopter lorsqu'il s'agit de définir le concept de « thème ». Parmi les nombreuses perspectives possibles, Rastier distingue entre autres l'approche lexicographique de l'approche sémantique. L'approche lexicographique, telle qu'esquissée par Rastier, s'apparente au thème en tant qu'« à-propos-de ». En effet, comme le souligne Rastier (2001, p. 196), « la voie lexicographique, tributaire d'une linguistique du signe, définit le thème, comme mot-vedette, généralement un substantif, auquel sont rapportés divers parasynonymes ou équivalents partiels : un dictionnaire de thème sera donc un sous-ensemble d'un dictionnaire. »

En contre partie, l'approche sémantique, davantage tributaire d'une linguistique du texte, ne réduit pas le thème exclusivement à un mot-vedette. Le thème est plutôt perçu au sein de « réseaux de récurrences et de transformations » (2001, p. 196). La perspective sémantique, transcendant la phrase au profit du texte, semble manifestement en adéquation avec certains des objectifs poursuivis dans le cadre de notre projet.

Dans la perspective de la sémantique interprétative, Rastier définit le thème dans les termes suivants : le thème est « une structure stable de traits sémantiques (ou sèmes), récurrente dans un corpus, et susceptible de lexicalisations diverses. » (2001, p. 197). L'opérationnalisation de cette définition présuppose donc que nous identifiions la nature de

ces traits sémantiques, du corpus, ainsi que des diverses lexicalisations possibles. Au sein des traits sémantiques, Rastier, en s'appuyant sur les travaux de Pottier (1974), oppose les sèmes génériques (un sème est un « élément d'un sémème, défini comme l'extrémité d'une relation fonctionnelle binaire entre sémèmes » (Rastier, 2001, p. 302)) aux sèmes spécifiques.

Les sèmes génériques sont ceux permettant d'indexer les sémèmes dans les classes sémantiques. Comme le notent Hébert (2001) et Rastier (1994), on distingue dans la sémantique interprétative quatre sortes de classes sémantiques auxquelles correspondent quatre sortes de thèmes génériques :

1) les thèmes taxémiques instanciés par les éléments d'un même taxème (i.e. classe de sémème minimale en langue (Rastier, 2001, p. 302)). Par exemple, le thème du //tabac// dans *Madame Bovary* se décline en « pipe », « cigarette » et « cigare » (Rastier, 1994, p.177);
2) les thèmes domaniaux instanciés par les éléments d'un même domaine sémantique;
3) les thèmes dimensionnels instanciés par les éléments d'une même dimension. Au sein des dimensions, on distinguera « les dimensions évaluatives ou thymiques, les tons, les espaces modaux, les plans temporels ou *chronotopes*. » (Rastier, 1994, p. 177);
4) les thèmes de champs instanciés par les éléments d'un même champ. « Les thèmes liés à un champ sémantique sont plus variables, puisque les champs ne sont pas des classes de langue. Les plus faciles à identifier sont ceux qui sont codifiés (topique littéraire) ou définitoires d'un corpus : par exemple, un corpus de notices d'entretien destinés au Falcon 900 se situe évidemment dans le domaine aéronautique (et les acceptions lexicales réfèreront toutes à ce domaine) mais elles en définissent un champ restreint. » (Rastier, 1994, p. 177-178).

En contre partie, les sèmes spécifiques sont à la base des thèmes spécifiques. Les thèmes spécifiques prennent la forme de regroupements récurrents de sèmes spécifiques. Il s'agit donc de molécules sémiques, lesquels sont définis en tant que « regroupement[s] stable[s] de sèmes, non nécessairement lexicalisé[s], ou dont la lexicalisation peut varier. » (Rastier, 1994, p. 223). Les thèmes peuvent donc faire l'objet d'expressions diverses allant du morphème au syntagme (que Rastier nomme, de manière générique, « lexicalisations »). Ces lexicalisations peuvent être de nature synthétique (manifestant alors au moins deux sèmes) ou de nature analytique (manifestant, dans ce cas, qu'un seul sème). À cet égard, Rastier soutient que la lexicalisation synthétique « ne jouit d'aucune prééminence théorique par rapport aux

autres lexicalisations : elle n'est pas le "mot juste" dont toutes les autres expressions ne seraient que des avatars. Selon les discours et les genres, les normes de lexicalisation des thèmes varient : la poésie lyrique cultive les lexicalisations analytiques, alors que dans les discours techniques, les synthétiques sont de rigueur. » (2001, p. 200). En outre, « un thème peut avoir une lexicalisation privilégiée (ex. : *ambition*), ou plusieurs (*pitié*, *commisération*, *compassion*). Il peut s'agir d'une lexie (*amour paternel*) ou n'avoir pas de nom retenu par l'usage (*sentiment du beau*, *amour de l'art*). » (Rastier, 2001, p. 200).

Dans cette perspective sémantique, un thème spécifique n'est rien d'autre qu'une molécule sémique. Le concept de « molécule sémique » est donc fondamental dans la poursuite de nos travaux, car c'est précisément sur la base de cette conception du thème que repose l'une des deux méthodologies visant à assister informatiquement l'identification du contenu thématique des documents. Rastier mentionne qu'un thème en tant que molécule sémique n'est pas nécessairement lexicalisé et, s'il l'est, celle-ci peut varier.

Cela nous amène à établir précisément la conception du thème interpellée dans nos processus visant à assister informatiquement l'identification des thèmes d'un corpus. Nous avons mentionné dans le chapitre précédent que notre projet consiste à explorer et à valider deux méthodologies informatiques (la première est fondée sur l'opération de catégorisation, alors que la seconde repose sur l'opération de classification). Dans une perspective d'assistance informatique à l'identification du contenu thématique fondée sur l'opération de catégorisation automatique, il importe que les thèmes soient lexicalisés afin de les traiter informatiquement. Dans la perspective de Rastier, il n'est cependant pas théoriquement nécessaire que les thèmes soient lexicalisés. Dans la sémantique interprétative, les thèmes spécifiques sont par principe indépendants de toute isotopie générique. Il n'y a donc pas *stricto sensu* de terme thématique propre à différents segments de documents, bien que le choix de certains termes thématiques soit, selon l'expression de Rastier, commode pour en représenter le contenu thématique.

La méthodologie fondée sur la classification automatique et sur l'extraction des termes thématiques candidats présuppose aussi une lexicalisation des thèmes. Comme nous le verrons, la méthodologie fondée sur la catégorisation s'inspire davantage d'une conception du thème en adéquation avec les travaux de Kintsch et Van Dijk, mais aussi de Rimmon-Kenan. En contrepartie, la seconde méthodologie (fondée sur la classification et l'extraction

automatique des termes thématiques) permet plutôt d'identifier des thèmes, spécifiques pour la plupart, dont la manifestation s'apparente étroitement à la perspective défendue par Rastier. Ainsi, si, dans la méthodologie fondée sur la catégorisation, le thème s'apparente à une étiquette générale représentant le contenu des documents auxquels elle est associée, il en va tout autrement dans la méthodologie classificatoire. Comme nous le verrons, nous défendons l'hypothèse selon laquelle, dans cette seconde démarche, les thèmes sont composés des lexicalisations de traits sémantiques caractéristiques – implicites dans certains cas – et récurrents dans les regroupements de segments de documents. Ne limitant pas les parcours thématiques possibles, en tant que parcours interprétatifs d'un corpus textuel, cette seconde méthodologie nous semble donc d'inspirer davantage de théorie de Rastier sur la thématique.

Nous avons établi le cadre théorique employé dans notre projet. En résumé, ce cadre théorique est caractérisé par les traits suivants.

1) En nous basant sur les travaux de Kintsch et Van Dijk, le concept de « thème » est abordé dans une perspective macrostructurelle (par opposition au thème phrastique, qui relève de la microstructure). Nous retenons aussi les trois opérations principales de la démarche de ces auteurs. Contrairement à ces derniers, nous ne les appliquons pas sur l'ensemble des propositions, mais plutôt sur l'ensemble des termes extraits de chacun des segments de texte préalablement regroupés.

2) L'inventaire des différentes conceptions du thème réalisé par Rimmon-Kenan nous permet de fonder et de spécifier davantage la perspective selon laquelle nous abordons la question du thème dans notre projet. À cet égard, nous partageons les principales caractéristiques de la conception qu'il propose. Selon cette dernière, certains thèmes (principalement ceux générés par l'opération de catégorisation) peuvent être définis en tant qu'étiquettes obtenues grâce à trois opérations : l'assemblage, la généralisation et l'étiquetage. Nous retenons en outre l'importance de la composante subjective du processus de thématisation.

3) Dans la perspective de la sémantique interprétative de Rastier, le thème est perçu au sein de « réseaux de récurrences et de transformations » (Rastier, 2001, p. 196). Il est défini en tant que « structure stable de traits sémantiques (ou sèmes), récurrente dans un corpus, et susceptible de lexicalisations diverses » (Rastier, 2001, p. 197). Au sein des traits

sémantiques, Rastier oppose les sèmes génériques aux sèmes spécifiques. Les sèmes génériques, auxquels correspondent les thèmes génériques, permettent d'indexer les sémèmes dans les classes sémantiques. Les sèmes spécifiques sont à la base des thèmes spécifiques. Selon Rastier, il s'agit de molécules sémiques, lesquelles sont définies en tant que « regroupement[s] stable[s] de sèmes, non nécessairement lexicalisé[s], ou dont la lexicalisation peut varier. » (Rastier, 1994, p. 223). Nous retenons donc que ces thèmes peuvent faire l'objet de diverses lexicalisations.

Dans les sections suivantes, nous présentons les différentes étapes des méthodologies proposées pour assister les opérations d'identification du contenu thématique et d'analyse thématique de documents textuels.

2.2. Méthodologie informatique[4]

2.2.1. Le prétraitement des documents

2.2.1.1. L'identification des unités et des domaines d'information

La première étape de notre démarche consiste à identifier, à partir des documents à analyser, les unités d'information (UNIFs) (i.e. les traits descriptifs qui serviront d'ancrage à l'analyse des segments de documents) et les domaines d'information (DOMIFs) (i.e. les segments de documents). Comme nous le verrons, ce sont ces deux objets issus du corpus (notre corpus d'expérimentation est décrit dans le chapitre 3) qui serviront d'éléments constitutifs de la matrice qui sera soumise aux modules de classification et de catégorisation. Cette étape est donc primordiale. Ces opérations reposent sur plusieurs enjeux théoriques issus principalement de l'analyse statistique et linguistique des données textuelles.

L'identification des unités d'information consiste à déterminer et à extraire les éléments sur la base desquels les différents segments du corpus à analyser seront comparés. Dans le cas des données textuelles, ces unités d'information peuvent prendre différentes formes. Elles

---

[4] La figure 2.16 (p. 77) présente de manière abrégée les principales étapes de la méthodologie informatique proposée.

peuvent être des mots, des mots composés, des phrases, des n-grammes[5], etc. Cette étape repose sur des décisions théoriques importantes. Comme le note Meunier (1995, p. 6), le fait de fonder une analyse des données textuelles sur des mots ou sur le lexique d'un corpus nous impose d'importants questionnements sur la nature même d'un mot et, de manière plus générale, des unités d'information présentes au sein d'un texte. S'agit-il uniquement d'une suite de caractères ? Un mot se définit-t-il par la séparation spatiale, l'identité morphologique, etc.? Ces questions fondamentales doivent nécessairement être abordées avant même d'envisager tout projet d'analyse et de gestion des documents textuels. Techniquement, cette opération, comme le note Memmi (2000), implique de soumettre les unités d'information à des considérations contradictoires, car elles se doivent d'être représentatives et discriminantes, tout en étant faciles à extraire. Il importe, en dernière instance, de se rappeler que toutes les décisions théoriques prises par le chercheur doivent prendre en compte les buts poursuivis, les résultats qu'il espère découvrir lors de ses recherches, car c'est à partir de ces éléments que la classification sera effectuée. Comme nous le verrons au troisième chapitre, dans le cadre de notre projet, les unités d'information retenues sont composées de l'ensemble des mots du lexique du corpus.

Suite à l'identification des unités d'information, il importe, par la suite, d'identifier les différents segments de textes qui seront comparés entre eux. Cette étape se fait sur la base du corpus à analyser. Encore une fois, ce processus n'est pas sans soulever plusieurs enjeux théoriques. De quelle nature doivent être les segments à comparer? Doivent-ils être identifiés sur la base de critères linguistiques? Est-il préférable de segmenter le corpus à analyser en pages (Moffat *et al.*, 1994)? En paragraphes? En phrases? Ou, en revanche, l'analyse doit-elle reposer sur des critères d'édition? Dans un tel cas, est-il préférable de segmenter selon certains marqueurs ou étiquettes (titre, chapitres, notes de bas de page, etc.) de type XML? Cette décision relève en dernière instance du chercheur et des données à analyser. Malgré les différentes possibilités qui s'offrent au chercheur, il semble préférable d'opter pour une segmentation qui n'est ni trop volumineuse (surtout lorsqu'il s'agit de documents très homogènes), ce qui donnerait lieu à une classification très grossière et à une perte

---

[5] Les *n*-grammes sont définis comme des séquences de *n* unités (très souvent des lettres). Ainsi, dans la phrase « Cette thèse est vraiment exceptionnelle. » se retrouve l'ensemble de tri-grammes suivants : {Cet, ett, tte, te_, e_t, ...}.

d'information importante; ni trop fine car, à l'inverse, les unités à comparer seraient alors beaucoup trop différentes les unes des autres pour les soumettre à toute forme de classification.

Comme nous l'avons mentionné précédemment, l'application d'un processus de segmentation des données à analyser constitue une particularité de notre démarche. En effet, les processus de classification et de catégorisation sont traditionnellement appliqués sur des documents entiers. Comme l'ont souligné très récemment Chali (2005) et Kim et Kim (2004), cette démarche, bien qu'efficace dans certains contextes, engendre un problème important. Dans cette optique il est, de fait, plutôt difficile d'attribuer automatiquement plusieurs étiquettes ou catégories thématiques à un document.

> *Most documents are about more than one subject, but many Natural Language Processing (NLP) and Information Retrieval (IR) techniques implicitly assume documents have just one topic. Even in the presence of a single topic within a document, the document may address multiple subtopics and various aspects of the primary topic. Hence, dividing documents into topically-coherent units and discovering their topic have many uses.* (Chali, 2005, p. 1)

Les recherches sur la thématique ont d'ailleurs démontré depuis longtemps l'importance que l'on doit accorder à la pluralité de thèmes présents dans un document, comme en témoigne cet extrait de Louwerse et Van Peer (2002, p. 3) à propos des écrits de Tomashevsky (1925) : « *According to Tomashevsky a text as a whole has a theme, built from smaller themes. [...] They all have a certain unity and are composed of even smaller thematic elements that are arranged in a definite order.* »

Afin de contourner cette limite, une solution « computationnellement » peu coûteuse consiste à effectuer d'abord une segmentation des documents, permettant ainsi d'attribuer facilement plusieurs catégories thématiques à chaque document du corpus (figure 2.1).

La segmentation des documents est actuellement un important sujet de recherche. Plusieurs méthodes de segmentation sont d'ailleurs explorées. Parmi les plus fréquemment citées, on retrouve celles fondées sur des marqueurs de discours (*discourse passages*) (Kaszkiel et Zobel, 2001; Callan, 1994; Hearst, 1994;) et sur les marqueurs sémantiques (*semantic passages*) (Kaszkiel et Zobel, 2001; Hearst, 1994). Dans le cadre de notre projet, nous privilégierons une segmentation fondée non pas sur certaines propriétés sémantiques des

documents (comme c'est le cas dans les deux méthodes mentionnées ci haut), mais plutôt sur des séquences ou suites de mots (*non-overlapping window passages*). Nous utiliserons une segmentation par paragraphe. Cette méthode possède de nombreux avantages. En effet, cette dernière est « computationnellement » peu coûteuse, tout en ne nécessitant aucune propriété structurelle explicite des documents (Kim et Kim, 2004).

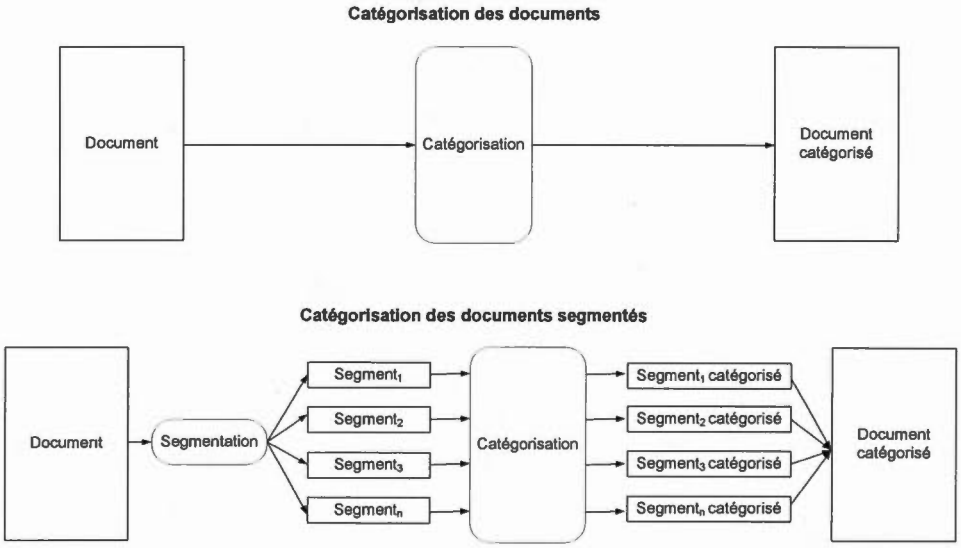


Figure 2.1. Comparaison des modèles de catégorisation des documents fondés sur les documents entiers et les documents segmentés (figure inspirée de Kim et Kim, 2004).

2.2.1.2. Le filtrage du lexique

L'étape suivante porte sur le lexique extrait du corpus à analyser. Il s'agit de lui appliquer différents filtres statistiques et linguistiques. La nature et l'importance de cette étape sont le centre de plusieurs débats théoriques. Ainsi, contrairement aux travaux de Riloff (1995), plusieurs recherches tendent à démontrer qu'un filtrage adéquat du lexique de départ permet d'éliminer plusieurs éléments susceptibles d'affecter l'« interprétabilité » des résultats, tout en diminuant substantiellement le temps nécessaire au traitement du corpus. Ainsi, comme l'indiquent Frakes et Baeza-Yates (1992, p. 113), « *it has been recognized since the earliest days of information retrieval (Luhn, 1957) that many of the most frequently occurring words in English (like "the", "of", "and", "to", etc.) are worthless indexing terms.* »

Bien que les propos de Frakes et Baeza-Yates portent sur l'indexation automatique de textes, nous sommes d'avis qu'ils s'appliquent également à d'autres tâches, telles la classification et la catégorisation. Comme nous le verrons dans les sections suivantes, le lexique (i.e. la liste des mots employés dans un corpus donné) constitue l'une des deux composantes principales de la matrice traitée par le classifieur et le catégoriseur; et la dimension de l'espace vectoriel dans lequel les domaines d'information seront comparés affecte directement le temps nécessaire à la classification de chacune des entrées de la matrice.

L'opération de filtrage du lexique est composée traditionnellement de plusieurs sous-opérations. La première d'entre elles consiste à supprimer certains termes non pertinents à l'analyse. Le fait de conserver ces termes, en plus d'affecter directement la qualité de la classification obtenue, ajoute aussi du bruit dans les résultats, ce qui a pour effet de compliquer le travail du chercheur lors de l'exploration et de la validation des classes produites[6]. Le filtrage du lexique peut être effectué à l'aide de plusieurs techniques, certaines étant de nature linguistique, d'autres de nature statistique. Une première opération a pour but de supprimer l'ensemble des mots fonctionnels présents dans le texte. Dans une perspective de traitement automatique des données textuelles, les mots fonctionnels ou les « mots vides » peuvent être définis comme étant l'ensemble des mots non pertinents à l'égard des buts poursuivis[7]. Pour plusieurs, ces mots fonctionnels prennent le nom de « *stop-words* » ou de

---

[6] Krause (1996, p. 6) note que, dans le domaine du repérage de l'information, le filtrage des mots fonctionnels permet de réduire le corpus de départ d'environ cinquante pourcent. Comme nous le verrons dans le quatrième chapitre, l'opération de filtrage du lexique que nous avons effectuée a permis de réduire la taille du lexique de plus de 90%.

[7] Selon Lallich-Boidin et Maret (2005), il est possible de définir une liste de mots vides selon deux perspectives. La première, d'inspiration linguistique, les définit comme étant les mots grammaticaux auxquels il est impossible d'associer un contenu sémantique spécifique. Dans cette perspective, les mots vides sont propres à chaque langue et ne varient pas en fonction des corpus et des traitements réalisés. La seconde perspective, davantage inspirée de la statistique textuelle, les définit comme étant les mots dont la fréquence est distribuée uniformément dans un corpus particulier. Ainsi, selon cette seconde perspective, une liste de mots vides est composée de l'ensemble des mots ne permettant pas de discriminer un texte parmi d'autres. Selon cette optique, les mots vides dépendent donc des objectifs à atteindre, mais aussi – et surtout – du corpus traité. Nous sommes partiellement en accord avec la position de Lallich-Boidin et Maret. D'un point de vue pratique, les mots vides sont effectivement non-pertinents à l'égard des buts poursuivis. Il ne nous semble cependant pas raisonnable de totalement réduire le critère linguistique d'identification des mots vides à la valeur discriminante de ces mêmes mots.

« *trivial-words* » (Popping, 2000). Ce processus est réalisé en retirant les termes du corpus figurant dans une liste prédéfinie de termes fonctionnels.

Par la suite, une deuxième opération est appliquée au lexique. Il s'agit cette fois-ci d'appliquer certains filtres statistiques au lexique du corpus afin d'en éliminer les termes qui, tout en ne figurant pas dans la liste des mots fonctionnels, ne sont pas pertinents à l'analyse. La pertinence des termes relève, dans le cadre de la classification et de la catégorisation des documents, de leur valeur en tant que facteurs discriminants. C'est sur la base de cette valeur que le système pourra juger du regroupement ou non de deux ou plusieurs segments de texte. Ainsi, il importe, afin d'optimiser les résultats obtenus, de supprimer les mots dont la fréquence est supérieure ou inférieure à un certain seuil (déterminé empiriquement). La figure 2.2 (tirée de Schultz, 1968, p. 120; adaptée dans Van Rijsbergen, 1979) représente bien le filtrage à effectuer d'un point de vue de la distribution statistique du lexique d'un corpus. La courbe de couleur rouge représente la distribution des mots du lexique par ordre décroissant. En ne tenant compte que de cette distribution, seuls les mots très fréquents peuvent être jugés représentatifs du corpus dans lequel ils figurent. La courbe pointillée de couleur bleue représente la valeur discriminante des termes en fonction de leur fréquence dans le corpus. Les deux traits verticaux de couleur noire représentent l'intervalle des termes retenus comme étant les plus discriminants ou représentatifs du corpus. En fonction de ces indications, on constate donc que les termes à retenir grâce aux différents processus de filtrage sont ceux que l'on retrouve dans la zone de couleur grise.

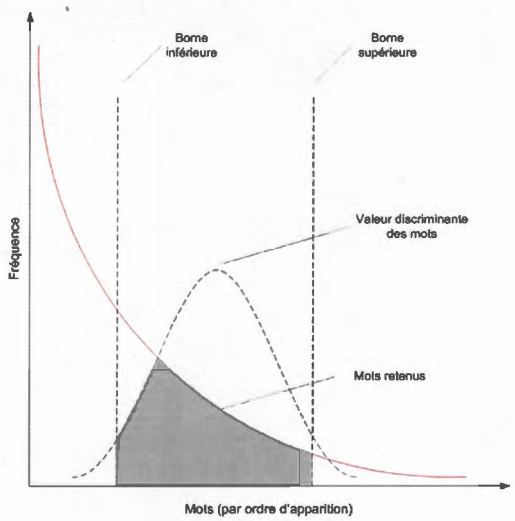


Figure 2.2. Les termes à retenir suite aux opérations statistiques de filtrage.

Une troisième étape s'impose pour le filtrage du lexique. Elle consiste à supprimer manuellement tous les termes non pertinents qui ont néanmoins résisté aux deux premières opérations. Finalement, une dernière étape de traitement du lexique est nécessaire afin d'optimiser le processus de classification. Aucun système de classification de nature exclusivement statistique n'est sensible aux variantes sémantiques et syntaxiques présentes dans un corpus; il importe alors d'appliquer au lexique du corpus une opération de lemmatisation. Il s'agit d'une phase importante du traitement. Elle se distingue, en outre, de l'opération de *stemming* qui est plus fréquemment employée dans le repérage de l'information. En effet, bien qu'il s'agisse de deux opérations comparables permettant de réduire le nombre de lexèmes présents dans un corpus, elles se distinguent par la manière dont elles procèdent et par les résultats qu'elles génèrent. L'opération de lemmatisation est réalisée généralement en effectuant d'abord un marquage morphosyntaxique des différents lexèmes à analyser, puis en les comparant à un dictionnaire. Ce premier processus permet de générer une liste de lemmes propres à une langue donnée. En revanche, le *stemming* (Porter, 1980) n'implique pas de marquage morphosyntaxique. Ce second processus est plutôt réalisé en appliquant un ensemble de règles aux lexèmes du corpus. Le résultat obtenu par le *stemming* est une liste de *stems* (racines). En fonction du marquage morphosyntaxique, l'opération de lemmatisation est beaucoup plus longue à réaliser, mais les résultats qu'elle génère sont souvent beaucoup plus détaillés et plus exacts que ceux générés par l'opération de *stemming*.

La lemmatisation est une opération très délicate, car elle implique un processus complexe de désambiguïsation sémantique. Les logiciels automatiques de désambiguïsation sémantique sont très souvent approximatifs et nécessitent en dernière instance une intervention humaine (Brunet, 2000, 2002). Comme le souligne Muller (1968, p. 143 *in* Lebart *et al.*, 1998, p. 24) concernant les normes nécessaires pour ce type d'opération :

> *The norm should be acceptable at the same time to linguists and their adjuncts as well as to statisticians. However their needs are often contradictory. Linguistic analysis results in loose classifications that always include indeterminate zones; the material it uses is eminently continuous, and only rarely is it possible to find clear boundaries; most of the time it requires a close inspection of the syntagmatic (contextual) and paradigmatic (lexical) surroundings before subdividing. On the other hand, in all its applications, statistics becomes involved in simplifying categories; it can only do its work once the continuity of language has been made discontinuous, which is harder at the lexical level than at other levels; it loses its effectiveness when distinctions increase or categories become smaller; when working on large data sets it has difficulty dealing with the type of thinking that prefers analyzing isolated facts.*

Malgré les nombreuses difficultés inhérentes à l'opération de lemmatisation, celle-ci demeure néanmoins une étape importante du processus de Lecture et d'Analyse de Textes Assistées par Ordinateur (LATAO). Dans notre projet, le prétraitement des documents sera effectué en employant certaines fonctionnalités de l'application commerciale de forage de textes WORDSTAT.

2.2.2. La vectorisation

La vectorisation repose sur les choix et les résultats des étapes précédentes. Il s'agit, à partir des résultats obtenus lors du processus de segmentation et de ceux obtenus lors de l'identification et du filtrage des unités d'information, de constituer une matrice formée, d'une part, des segments de textes et, d'autre part, des unités d'information (figure 2.3). Autrement dit, le texte est traduit dans un modèle vectoriel (Salton, 1989). L'objectif est donc de parvenir à une représentation matricielle du texte de départ composée des vecteurs $\vec{\xi}$ représentant chacun des segments ou domaines d'information. Ces vecteurs peuvent être représentés de la manière suivante $\vec{\xi} = \{0,1\}^N$; où N est le nombre total de mots différents dans le corpus filtré et correspond à la dimension de l'espace vectoriel dans lequel le texte à traiter est représenté.

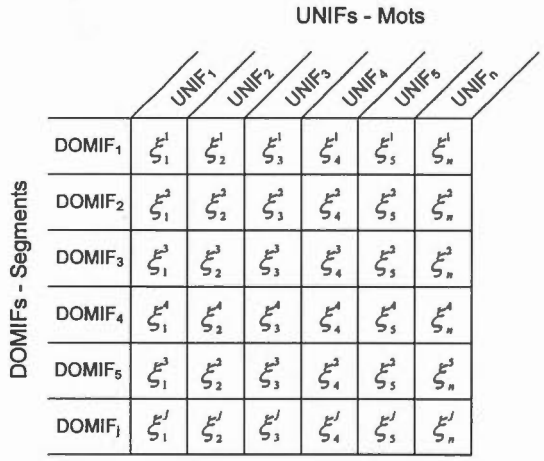


Figure 2.3. La matrice composée des domaines d'informations (DOMIFs) (segments) et des unités d'information (UNIFs) (termes).

Les entrées de la matrice varient en fonction du critère sélectionné (absence, présence, poids, etc.). Comme nous mènerons deux expérimentations distinctes, nous générerons deux matrices. La première sera soumise en intrant au module de catégorisation. À l'étape de catégorisation, comme nous emploierons un réseau hybride neuro-flou, la matrice devra être pondérée. Le critère de pondération employée est celui de la fréquence d'apparition des termes dans chacun des segments (paragraphes). La seconde matrice sera soumise en intrant au module de classification. Comme nous utiliserons le classifieur ART1 à cette étape de notre expérimentation, nous opterons pour une matrice binaire[8]. Les paramètres utilisés pour créer cette matrice sont essentiellement de nature statistique, c'est-à-dire que l'on évalue l'occurrence (pondérée, en ce qui concerne la matrice soumise à l'opération de catégorisation; binaire, en ce qui concerne la matrice soumise à l'opération de classification) de chaque mot (unité d'information) du texte en fonction de sa présence dans chacun des segments (domaines d'information). Plusieurs types de classifieurs et de catégoriseurs ont été explorés avec succès sur ce type de matrice. Ils vont de l'Analyse en Composante Principale

[8] Il est à mentionner que ces deux matrices seront « creuses » (*sparse*), car elles seront composées principalement de composantes nulles. De plus, comme les vecteurs qui les composent auront de nombreuses dimensions (plusieurs dizaines) cela pose un important problème de calcul et de représentation. Comme le mentionne Memmi (2000), c'est ce que l'on nomme communément la « malédiction dimensionnelle » (*curse of dimensionality*). C'est donc ici que le filtrage des unités d'information prend toute son importance.

(ACP) (Lebart et Salem, 1988) aux k-moyens (Balpe, Lelu et Papy, 1996), en passant par les réseaux de neurones (Kohonen, 2001 ; Nault, Rialle et Meunier, 1999; Salton, Buckley et Allan, 1994; Veronis *et al.*, 1990) et les algorithmes génétiques (Nault, 2000).

Dans la figure 2, la matrice composée des domaines d'information (paragraphes) et des unités d'information (mots) est de taille (n x j). Elle est construite en indiquant l'absence ou la présence de chacune des mots dans chacun des paragraphes. Ainsi, chaque paragraphe est représenté sous forme de vecteurs. Les classifieurs, ainsi que les catégoriseurs utilisent donc l'ensemble d'intrants $\vec{\xi}^{\mu} = \{0,1\}^N$ où μ est représenté par le nombre de segments.

2.2.3. Le regroupement et l'identification du contenu des documents

Dans notre projet, nous avons comme objectif d'explorer la pertinence de deux démarches méthodologiques. La troisième étape du processus se divise donc en deux opérations possibles, à savoir, d'une part, la catégorisation automatique et, d'autre part, la classification automatique.

Ces deux opérations ont été l'objet de plusieurs travaux dans les domaines de l'intelligence artificielle et de l'apprentissage machine. Dans ces domaines, elles s'inscrivent dans deux catégories de techniques informatiques distinctes. D'abord, la catégorisation automatique figure dans la catégorie des méthodes dites supervisées. De manière générale, les méthodes supervisées impliquent une intervention directe de l'utilisateur dans le processus réalisé. Cette intervention peut prendre différentes formes. Dans le cas de la catégorisation automatique, cette intervention consiste à guider le processus d'apprentissage en indiquant d'abord au système quelles catégories doivent être associées à certains documents. L'ensemble de ces documents catégorisés manuellement porte le nom d'ensemble d'apprentissage. Sur la base de cet ensemble d'apprentissage, l'algorithme de catégorisation effectue un apprentissage (la nature de celui-ci est fonction de la nature de l'algorithme employé), puis projette l'information apprise sur un ensemble de test constitué de documents dont le système ne connaît pas *a priori* les catégories auxquelles ils appartiennent.

En contrepartie, la classification automatique figure traditionnellement dans la catégorie des méthodes dites non-supervisées. Les différentes méthodes figurant dans cette catégorie

sont réalisées de manière totalement autonome. Elles sont donc réalisées sans aucune intervention de l'utilisateur. Tout comme les méthodes supervisées, les méthodes non-supervisées intègrent aussi un mécanisme d'apprentissage, mais ce dernier opère uniquement sur la base de l'information soumise au système, sans avoir recours à des informations ou à des métadonnées devant être spécifiées par l'utilisateur. Dans le cas de la classification automatique, l'algorithme utilisé peut aussi nécessiter l'utilisation d'un ensemble d'apprentissage et d'un ensemble de test, mais il n'est pas nécessaire qu'il en soit ainsi. L'objectif de l'opération de classification consiste à regrouper les différentes données dans des regroupements les plus homogènes possible en employant uniquement les traits caractéristiques ayant servi à décrire chaque élément.

La distinction entre ces deux opérations se manifeste aussi lorsque ces opérations sont appliquées dans les domaines du Traitement Automatique du Langage (TAL) et de la Gestion Électronique des Documents (GÉD). En effet, ces deux opérations correspondent à deux types d'analyses distincts. L'opération de catégorisation automatique, étant réalisée en employant certaines métadonnées préalablement connues, est très étroitement reliée à des processus d'analyse de nature prédictive ou normative. Par exemple, une telle opération semble des plus adaptée à l'identification d'auteur (*authorship attribution*) (Diederich *et al.*, 2000 ; Stamatatos, Kokkinakis et Fakotakis, 2000) et à l'identification du contenu thématique des documents.

L'opération de catégorisation présuppose cependant que nous disposions *a priori* d'informations valides ou attestées pouvant être employées pour effectuer l'apprentissage. En ce sens, au niveau théorique, l'opération de catégorisation automatique interpelle nécessairement des informations externes aux données à traiter. En d'autres termes, la réalisation de cette opération implique l'application d'une structure d'informations sur les données textuelles. Dans une tâche de catégorisation thématique, cette structure d'informations externe prend la forme d'une taxinomie de catégories thématiques ou d'un plan de classification.

En contre partie, l'opération de classification automatique est effectuée en ne considérant que les données provenant exclusivement des données textuelles (i.e. le corpus) soumises au système. Elle ne consiste pas à appliquer une structure d'informations externes (i.e. métadonnées) aux documents. L'opération de classification vise à regrouper des documents

(ou des segments de documents) en ne considérant que leur contenu. Il s'agit d'une opération exploratoire et descriptive. L'opération de classification possède donc un niveau d'objectivité supérieur à celui de la catégorisation automatique.

2.2.3.1. La catégorisation thématique

La première démarche méthodologique, de nature prédictive, repose sur l'application d'une opération de catégorisation automatique des segments (paragraphes). Dans cette démarche, les vecteurs soumis au processus de catégorisation sont pondérés en fonction de la fréquence d'apparition de chaque mot dans chaque segment. D'un point de vue informatique, la catégorisation des documents est un processus très différent de celui de la classification. Contrairement au processus de classification (processus décrit à la section 2.4.2) dont l'objectif consiste à regrouper des documents partageant certains critères de similarité internes (à ces mêmes documents), la catégorisation des documents est définie comme un processus d'organisation supervisé dans le cadre duquel une ou plusieurs catégories thématiques (il s'agit donc de critères de similarité externes aux documents) sont attribuées à chacun des documents (Sebastiani, 2005b, 2005a; Manning et Schütze, 1999;). En d'autres termes, cette tâche réside dans la projection d'une taxinomie (c'est-à-dire un ensemble de catégories structurées) sur des documents afin d'attribuer à chaque document une ou plusieurs étiquettes thématiques représentant le contenu de chacun de ces documents.

Le processus de catégorisation des documents peut être formalisé dans les termes suivants (Sebastiani, 2002). Il s'agit d'une fonction d'appariement $\Phi : D \times C \rightarrow \{T, F\}$ où C est une liste prédéfinie de catégories $\{c_1, c_2, ..., c_n\}$ et D est un ensemble de documents $\{d_1, d_2, ..., d_n\}$. Le processus de catégorisation peut consister à attribuer une seule catégorie par document (auquel cas $c_i \in C \rightarrow d_j \in D$) ou plusieurs catégories par documents (auquel cas $0 < n_j \leq |C| \rightarrow d_j \in D$).

Dans son application au traitement automatique des documents, le processus de catégorisation implique plusieurs considérations. Premièrement, le processus de catégorisation présuppose l'élaboration *a priori* d'une taxinomie ou d'une hiérarchie de catégories thématiques adaptées au contenu et aux spécificités des documents à traiter. Ce processus implique aussi l'identification, à partir de l'ensemble des documents, des

caractéristiques (linguistiques et statistiques) qui serviront de base à la catégorisation. De plus, ce processus implique le choix d'une méthode d'attribution des catégories aux différents documents. Traditionnellement, dans le domaine de l'analyse de contenu, l'assignation des catégories est effectuée manuellement. C'est l'analyste qui, en fonction de son propre répertoire de catégories, attribue les catégories à assigner à chaque document. Cependant, grâce aux récentes recherches dans les domaines de l'apprentissage machine, il est possible d'assister informatiquement ce processus. À cet égard, plusieurs méthodes ont été explorées. Les plus fréquemment employées sont les réseaux de neurones artificiels, les algorithmes de décision (tel l'algorithme C4.5 (Quinlan, 1993)) et d'induction de règles (tel l'algorithme CN2, (Clark et Niblett, 1989)). Actuellement, un important effort de recherche est consacré à l'exploration des *Support Vector Machine* dans leur application à la catégorisation de documents (Basu, Watters, et Shepherd, 2003; Joachims, 2002).

Comme en témoigne la littérature dans le domaine du repérage de l'information et du Traitement Automatique du Langage (TAL), le perceptron multicouches est un réseau des plus adaptées pour la catégorisation des documents (De Pasquale et Meunier, 2003; Ruiz et Srinivasan, 1998). Malgré leurs avantages (robustesse, apprentissage, rapidité, etc.), les différentes méthodes connexionnistes, dans leur application à la catégorisation automatique, possèdent plusieurs inconvénients importants. En effet, elles sont souvent perçues comme des « boîtes noires » ne permettant généralement pas d'obtenir d'informations explicites sur la dynamique interne du processus. En outre, elles ne permettent pas d'extraire des règles (qui seraient, manifestement, très utiles pour comprendre le mécanisme de catégorisation), ne permettent pas l'intégration de connaissances *a priori* et ne peuvent garantir la convergence des résultats. Compte tenu des limites inhérentes aux approches connexionnistes, nous avons privilégié dans notre projet une approche hybride neuro-floue pour effectuer la catégorisation des segments de documents. L'application informatique que nous avons employée pour réaliser l'opération de catégorisation est NEFCLASS-J.

2.2.3.1.1. NEFCLASS-J

NEFCLASS-J (Nauck, 1999) est une application hybride programmée en Java combinant une approche neuronale avec certains principes de la logique floue (Zadeh, 1995) afin de

catégoriser des données. Comme nous l'avons mentionné, le recours à diverses techniques neuronales pour la catégorisation des données est une pratique courante dans de nombreux domaines de recherche (De Pasquale et Meunier, 2003; Forest, 2002; Sebastiani, 2002; Manning et Schütze, 1999; Yang, 1999; Ruiz et Srinivasan, 1998). Cependant, l'introduction d'une composante floue au sein de ces techniques fait encore l'objet de plusieurs recherches importantes (Nauck, 1999). Ces travaux sont principalement motivés par le besoin de fournir à l'utilisateur de l'information facilement interprétable concernant les différentes opérations ayant mené aux résultats de la catégorisation. Ainsi, l'objectif de combiner les réseaux de neurones et la logique floue consiste précisément dans le développement de catégoriseurs interprétables (Nauck, 1999).

L'application NEFCLASS-J s'insère précisément dans cette optique. Il s'agit d'un système hybride neuro-flou composé de deux modules principaux, l'un intégrant des principes de la logique floue (i.e. un contrôleur flou); l'autre composé d'un réseau de neurones de type perceptron multicouches. Un tel système hybride est caractérisé par les traits suivants (Nauck, 1999, p. 18) : 1) l'apprentissage est effectué par un algorithme dérivé d'un réseau de neurones, 2) l'architecture générale de ce système peut être représentée par un réseau récurrent (mais il n'est pas nécessaire qu'il en soit ainsi), 3) ce système peut toujours être interprété en termes de règles de la forme « si... alors... », 4) l'apprentissage est effectué à partir de la sémantique du modèle flou sous-jacent, permettant ainsi de préserver l'interprétabilité linguistique du modèle et 5) ce modèle effectue une approximation de fonction.

Les systèmes neuro-flous se distinguent ainsi des réseaux de neurones flous (comme FUZZY ART ou FUZZY ARTMAP). Nauck (1999, p. 18) présente clairement cette distinction dans les termes suivants :

> *On the other hand a fuzzy neural network is a neural network that uses fuzzy methods to learn faster or perform better. In this case the improvement of the neural network is the main intention. An interpretation in terms of fuzzy rules is neither important nor possible here, because the system is based on a neural network with black box characteristics.*

Comme nous le verrons, le recours à une telle approche neuro-floue permet de contourner les inconvénients inhérents à chacune des approches, tout en en conservant leurs avantages respectifs. Comme le souligne Nauck (1999, p. 18) :

Neuro-fuzzy systems are created to overcome the disadvantages of neural networks and fuzzy systems. The term is usually used for every kind of combination of neural networks and fuzzy systems. One approach is to combine both in such a way that learning algotithms (sic) *are used to determine parameters of fuzzy systems. This means that the main intention of a neuro-fuzzy approach is to create or improve a fuzzy system automatically by means of neural network methods. An even more important aspect is that the system should always be interpretable in terms of fuzzy if-then rules, because it is based on a fuzzy system reflecting vague knowledge. In a word: the task is to overcome the disadvantages without loosing* (sic) *the advantages.*

2.2.3.1.1.1. Le module flou

Le module flou de NEFCLASS-J est à la base de deux principales fonctionnalités du système : i) l'apprentissage des règles floues et ii) la génération des ensembles flous. Ceci signifie que l'application permet d'identifier des règles floues dont la forme est la suivante :

if [CONDITION 1] MOT_1 *is « fréquence floue »* and
if [CONDITION 2] MOT_2 *is « fréquence floue »* and
if [CONDITION 3] MOT_3 *is « fréquence floue »* and
if [CONDITION N] MOT_N *is « fréquence floue »*
then [CONCLUSION] **CATÉGORIE**$_X$

Voici un exemple de règle (non-élaguée) automatiquement extraite à partir de notre corpus d'expérimentation :

if *ARABE is small* [CONDITION 1] and *CINEMA is small* [CONDITION 2] and *CLIJSTERS is small* [CONDITION 3] and *COLLEGE is small* [CONDITION 4] and *COMMISSION is small* [CONDITION 5] and *CONSEIL is small* [CONDITION 5] and *CONSOMMATEUR is small* [CONDITION 6] and *CREATION is small* [CONDITION 7] and *CUISINE is small* [CONDITION 8] and *DIRECTEUR is small* [CONDITION 9] and *DOCTEUR is small* [CONDITION 10] and *DROIT is small* [CONDITION 11] and *ELECTRABEL is small* [CONDITION 12] and *EUROS is small* [CONDITION 13] and *FER is small* [CONDITION 14] and *FILM is small* [CONDITION 15] and *FOI is small* [CONDITION 16] and *FORMATION is small* [CONDITION 17] and *FROMAGER is small* [CONDITION 18] and *GARE is small* [CONDITION 19] and *GOUVERNEMENT is small* [CONDITION 20] and *GUERRE is small* [CONDITION 20] and *HENIN is small* [CONDITION 21] and *INFORMATIQUE is small* [CONDITION 22] and *IRAK is small* [CONDITION 23] and ***ISLAMIQUE is medium*** [CONDITION 24] and *JUGE is small* [CONDITION 25] and *JUSTICE is small* [CONDITION 26] and *LABORATOIRE is small* [CONDITION 27] and *LOGICIEL is small* [CONDITION 28] and *LOI is small* [CONDITION 29] and *MALADE is small* [CONDITION 30] and *MASTERS is small* [CONDITION 31] and *MATCH is small* [CONDITION 32] and *MEDECIN is small* [CONDITION 33] and *MEDECINE is small* [CONDITION 34] and *MEDICAL is small* [CONDITION 35] and *NUCLEAIRE is small* [CONDITION 36] and *NUMERIQUE is small* [CONDITION 37] and *ORDINATEUR is small* [CONDITION 38] and *PEINE is small* [CONDITION 39] and *PLAINTE is small* [CONDITION 40] and *POLICE is small* [CONDITION 41] and *POLITIQUE is small* [CONDITION 42] and *POMME is small* [CONDITION 43] and *PRISON is small* [CONDITION 44] and *PROCUREUR is small* [CONDITION 45] and *PRODUCTEUR is small* [CONDITION 46] and *PROFESSEUR is small* [CONDITION 47] and *PROFESSIONNEL is small* [CONDITION 48] and *RECETTE is small* [CONDITION 49] and ***RELIGIEUX is large*** [CONDITION 50] and *REALISATEUR is small* [CONDITION 51] and *RESEAU is small* [CONDITION 52] and *SANTE is small* [CONDITION 53] and *SCIENTIFIQUE is small* [CONDITION 54] and *SET is small* [CONDITION 55] and *SNCB is small* [CONDITION 56] and *SPORT is small* [CONDITION 57] and *SPORTIF is small* [CONDITION 58] and *SYSTEME is small* [CONDITION 59] and *TECHNOLOGIE is small* [CONDITION 60] and *TENNIS is small* [CONDITION 61] and *TERRAIN is small* [CONDITION 62] and *TOURNOI is small* [CONDITION 63] and *TRAIN is small* [CONDITION 64] and *TRIBUNAL is small* [CONDITION 65] and *UNIVERSITAIRE is small* [CONDITION 66] and *UNIVERSITE is small* [CONDITION 67] and *VAINQUEUR is small* [CONDITION 68] and *VICTIME is small* [CONDITION 69] and *VICTOIRE is small* [CONDITION 70] and *VIN is small* [CONDITION 71] and *VIOLENCE is small* [CONDITION 72] and *VELO is small* [CONDITION 73] and *ELECTRIQUE is small* [CONDITION 74] and *ELEVE is small* [CONDITION 75] and *ETUDE is small* [CONDITION 76] and *ETUDIER is small* [CONDITION 77] then **ISLAM [CONCLUSION]**

L'identification de ces règles est d'abord réalisée en fuzzifiant automatiquement la présence de chacun des termes dans chacun des documents (voir figure 2.4). Par la suite, le système identifie les règles optimales permettant de catégoriser les documents. Les règles extraites sont donc directement impliquées dans le processus de catégorisation. C'est sur la base de ces règles automatiquement extraites que les documents sont catégorisés.

Les règles floues sont automatiquement apprises (c'est-à-dire automatiquement extraites et archivées dans une base de connaissances) par le système à partir de l'ensemble de données qui lui sont fournies. Cependant, il est aussi possible (mais non nécessaire) pour l'utilisateur d'intégrer manuellement certaines connaissances dans le système. L'insertion de connaissances peut être effectuée soit en modifiant certaines règles apprises par le système, soit en en ajoutant ou en en supprimant.

Une fonctionnalité essentielle de l'application réside dans la possibilité d'effectuer un élagage automatique des règles apprises par le système. L'élagage permet de faciliter l'interprétation des règles, tout en en diminuant très souvent le nombre (l'exemple précédent illustre très bien la complexité que peuvent avoir certaines règles). Comme l'explique Nauck (Nauck, 1999, p. 31-32), le processus d'élagage est effectué en utilisant deux stratégies. La première consiste à supprimer des termes (c'est-à-dire des variables (*features*)) de l'antécédent de certaines règles. Cette stratégie est employée dans deux contextes spécifiques : elle sera employée, d'une part, lorsque certaines variables (rappelons que, dans le cadre de ce projet, les variables du système sont des termes) sont jugées par le système comme étant non-pertinentes à des fins de classification et, d'autre part, lorsque les termes employés dans des ensembles flous possèdent un coefficient de support très élevé. La seconde stratégie consiste à supprimer les règles non significatives ne contribuant que très peu aux performances du processus de catégorisation. Le processus d'élagage est des plus efficaces, car, en plus de faciliter l'interprétation des règles, il permet d'augmenter significativement les performances prédictives du système.

Par ailleurs, les ensembles flous servent de base à la génération automatique des règles floues. Les ensembles flous représentent les différents degrés que peut prendre chacune des variables indiquées dans l'ensemble des données soumises au système. Par exemple, la traduction en valeurs floues de la fréquence d'une variable particulière prendra la forme suivante (figure 2.4) :

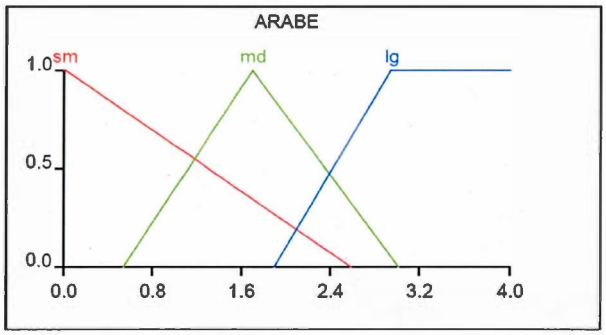


Figure 2.4. Fuzzification de la fréquence de la variable d'entrée « Arabe ».

Dans cette figure, on constate que la fréquence du mot « Arabe » prendra la valeur floue « faible » (*small* (sm)) si elle se situe entre 0 et 2,7; elle prendra la valeur floue « moyen » (*medium* (md)) si elle se situe entre 0,6 et 3 et elle prendra la valeur floue « élevé » (*large* (lg)) si elle se situe entre 1,8 et 4. Comme nous l'avons mentionné, ce processus de fuzzification est effectué de manière totalement automatique dans l'application NEFCLASS-J.

Ainsi, en utilisant ces deux mécanismes, l'application NEFCLASS-J permet de solutionner des problèmes de catégorisation de données en intégrant et en traitant de manière rigoureuse des connaissances vagues et en permettant l'attribution de degrés d'appartenance à chaque classe pour chacun des éléments à catégoriser.

2.2.3.1.1.2. Le perceptron multicouches

Le second module de l'application NEFCLASS-J est composé d'un réseau de neurones de type perceptron multicouches (Rosenblatt, 1958). Ce type de réseau est composé d'une couche d'entrée, d'une ou de plusieurs couches cachées et d'une couche de sortie (voir figure 2.5).

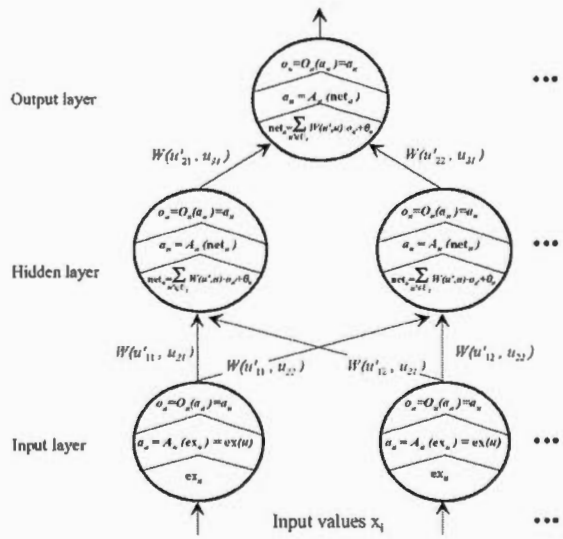


Figure 2.5. L'architecture générale du perceptron multicouches (tiré de Nauck, 1999, p. 14).

Dans l'application NEFCLASS-J, le perceptron multicouches est paramétré selon les spécifications suivantes. Au niveau architectural, le réseau est composé d'une couche d'entrée, d'une couche cachée et d'une couche de sortie. Le nombre de neurones de la couche d'entrée correspond au nombre de variables présentes dans les patrons d'apprentissage. La couche cachée correspond aux règles floues apprises par le système. Le nombre de neurones de cette couche est équivalent au nombre de règles. La couche de sortie correspond aux variables de sortie. Il y a donc dans cette troisième couche autant de neurones que de catégories possibles. Au niveau des paramètres d'apprentissage, les fonctions d'activation utilisées sont la norme-t (*t-norm*) et la conorme-t (*t-conorm*). Les ensembles flous sont représentés par les poids de connexion (flous) entre les neurones. Finalement, l'apprentissage est réalisé par rétropropagation. À partir de ces informations, l'architecture générale de l'application NEFCLASS-J peut donc être représentée ainsi (figure 2.6) :

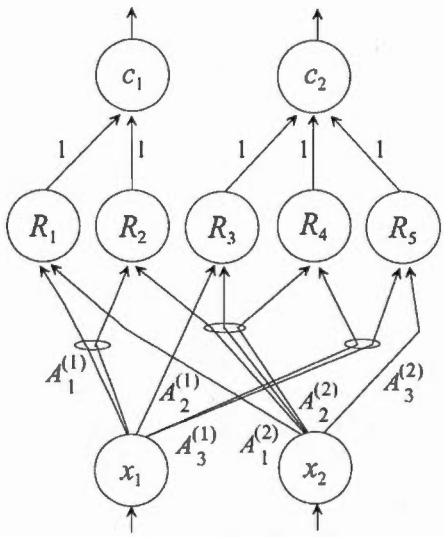


Figure 2.6. L'architecture générale de l'application NEFCLASS-J (tiré de Nauck, 1999, p. 23).

2.2.3.1.1.3. Avantages et inconvénients découlant de l'utilisation de la logique floue à des fins de catégorisation

L'utilisation de techniques hybrides neuro-floues à des fins de catégorisation de textes possède de nombreux avantages. En plus de fournir des résultats très impressionnants (Forest, 2005a, 2005b), une approche hybride neuro-floue permet d'obtenir plusieurs informations utiles afin d'interpréter les résultats : des ensembles flous pour chacune des variables et un ensemble de règles nous permettant de mieux comprendre les éléments ayant servi à effectuer le processus de catégorisation. À titre d'exemple, lors d'expérimentations préliminaires menées sur un échantillon de notre corpus composé de cent articles préalablement catégorisés manuellement en deux catégories (« TENNIS » et « GASTRONOMIE »), voici les deux règles (obtenues suite au processus d'élagage) permettant d'expliquer le processus ayant mené à la catégorisation du corpus : 1) IF « Tennis » (en tant que variable) *is small* [CONDITION 1] THEN « GASTRONOMIE » (en tant que catégorie) [CONCLUSION] et 2) IF « Tennis » (en tant que variable) *is large* [CONDITION 1] THEN « TENNIS » (en tant que catégorie) [CONCLUSION]. À partir de ces deux règles extraites, nous sommes donc en mesure d'affirmer que la variable « Tennis » peut permettre à elle seule de catégoriser correctement l'échantillon de corpus sur lequel nous nous sommes attardés. Ainsi, si cette variable est

fortement présente dans un document, alors il s'agit indubitablement d'un document appartenant à la catégorie « TENNIS ». En contrepartie, si cette même variable n'y est que faiblement présente, alors le document doit être catégorisé dans la catégorie « GASTRONOMIE ». Bien qu'évidentes, ces informations sont très utiles, car elles permettent à l'utilisateur d'identifier les variables importantes sur la base desquelles la catégorisation repose. Ce type d'informations n'est malheureusement pas disponible lorsque le processus de catégorisation est effectué par un simple réseau de neurones de type perceptron multicouches.

En contrepartie, le chercheur qui utilisera néanmoins un réseau de neurones de type perceptron multicouches (sans aucun mécanisme de logique floue) bénéficiera à tout le moins d'un avantage non négligeable : un temps de traitement plus rapide. En effet, le temps de traitement pour effectuer la tâche de catégorisation s'avère généralement entre deux et trois fois plus court avec un perceptron multicouches qu'avec un système hydride neuro-flou. Le tableau (2.1) suivant présente de manière synthétique les principaux avantages d'un système hybride neuro-flou (par opposition à un réseau de neurones).

Réseaux neuronaux	**Systèmes à base de logique floue**
Avantages	
• Formalisme robuste • Plusieurs architectures possibles • Plusieurs algorithmes d'apprentissage possibles • Rapide et efficace • Permet de traiter des données bruitées	• Formalisme robuste • Ne nécessite aucun modèle mathématique • Permet l'intégration de connaissances *a priori* sous forme de règles • Suivi et interprétation des processus et des résultats • Implémentation relativement facile • Permet l'extraction de règles
Inconvénients	
• Aucune information explicite sur la dynamique interne (boîte noire) • Ne permet pas l'extraction de règles • Requiert possiblement des phases de réapprentissage • Ne permet pas l'intégration de connaissances *a priori* • Convergence incertaine des résultats	• Aucun apprentissage possible • Adaptation relativement difficile aux modifications des données • Calibrage un peu difficile et incertain • Nécessite un processus de dé-fuzzification • Sensible aux données bruitées

Tableau 2.1. Les principaux avantages et inconvénients des réseaux de neurones et des systèmes à base de logique floue (tiré de Nauck, 1999, p. 18).

2.2.3.2. La classification

La seconde méthodologie explorée est de nature exploratoire. Elle est fondée sur une opération de classification des segments de documents. Dans le domaine du repérage de l'information, la classification des documents est une opération non supervisée visant à regrouper un ensemble de documents sur la base d'un ou de plusieurs critères de similarité (Jain *et al.*, 1999; Manning et Schütze, 1999). Traditionnellement, cette tâche de regroupement porte le nom de *clustering*. Ainsi, comme le soulignent Baeza-Yates et Ribeiro-Neto (1999, p. 438), « [Clustering is] *the grouping of documents which satisfies a set of common properties. The aim is to assemble together documents which are related among themselves.* »

Formellement, le processus de classification peut être perçu comme un système abstrait dont l'objectif est de regrouper des documents entre eux, en fonction de certains critères de similarité. En termes logiques, il s'agit d'une opération qui peut être définie de la manière suivante (Meunier, Remaki et Forest, 1999) : soit le quadruplet (O, X, I, G) où O est un ensemble d'objets (o_1... o_n); X est l'ensemble des caractéristiques (x_1.... x_n) décrivant chaque objet O; I est l'ensemble des types (i_1.... i_n); G est une fonction discriminante. À partir de ces informations, une opération de classification est définie ainsi : pour tout objet de l'ensemble O, (($G(x_1$.... $x_n)i)i_j$). En d'autres termes, la classification est une opération G qui, prenant en intrant un ensemble d'objets de type I décrits par leurs caractéristiques X, génère d'autres objets de type I_j. De façon plus générale, la classification est donc une opération abstraite (ou une fonction d'équivalence appliquée sur les objets d'un domaine) réalisant des classements ou des regroupements d'objets. Elle peut donc aussi être définie comme la projection d'une partition sur des objets, de façon à ce que ces objets soient regroupés de la manière la plus homogène possible.

Comme le souligne Charniak (1993, p. 135), le défi de la classification s'organise autour de trois axes principaux. Le premier, auquel nous avons fait allusion dans les sections précédentes, concerne la description des différents éléments à soumettre à l'opération de classification. En effet, il importe, à cet égard, de ne retenir que les traits caractéristiques (*features*) les plus discriminants, sur la base desquels les différents segments de documents seront comparés. Très souvent, cette description implique l'application sur les données d'une

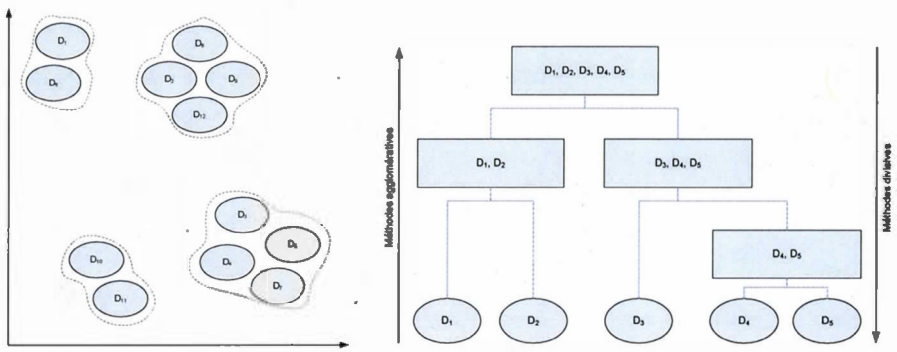
ou de plusieurs fonctions de filtrage et de nettoyage. En effet, pour obtenir une classification pertinente, il importe que le processus de classification repose sur des descripteurs significatifs des objets. Les décisions prises à cet égard doivent tenir compte autant des objectifs de la classification que de la nature même des objets à classer. Le second défi concerne la fonction discriminante. Cette fonction est à la base de la classification. Plusieurs hypothèses théoriques doivent être confrontées lorsqu'il s'agit de choisir la fonction discriminante la plus adaptée aux objectifs à atteindre. Cette fonction discriminante peut faire appel à plusieurs critères : l'identité, la similarité, l'homogénéité, l'équivalence, etc. Finalement, le dernier défi concerne le choix de l'algorithme employé pour effectuer l'opération de classification. Ce choix implique des enjeux tant théoriques que pratiques. En effet, compte tenu des récents travaux dans le domaine de la classification des données, plusieurs possibilités sont ouvertes. Le choix d'une approche au détriment d'une autre fait intervenir des considérations concernant tant la nature des données à regrouper que les caractéristiques de la classification souhaitée (exclusivité, hiérarchie, dynamicité et incrémentalité, etc.).

2.2.3.2.1. Quelques techniques classiques pour la classification automatique des données textuelles

Plusieurs techniques informatiques ont été explorées afin d'accomplir des tâches de classification des données textuelles (Jain *et al.* 1999; Manning et Schütze, 1999). Il est possible de regrouper l'ensemble des techniques et des méthodes de classification de différentes manières. Certains distinguent les différentes méthodes de classification en fonction de la rigidité des regroupements effectués (*hard clustering* vs *soft clustering*) (Manning et Schütze, 1999). Les méthodes de *hard clustering* permettent de positionner chaque objet dans un seul (et seulement un seul) regroupement. Par opposition, certaines techniques de *soft clustering* permettent de situer les objets dans un ou plusieurs regroupements en attribuant à chaque objet un degré d'appartenance à chacun des regroupements. Certaines techniques de regroupement intégrant des principes de la logique floue font partie de cette seconde catégorie.

Par ailleurs, les techniques de *hard clustering* peuvent aussi faire l'objet de distinctions supplémentaires. En effet, au sein de cette catégorie, il est possible de distinguer, d'une part, les

techniques effectuant des regroupements ou des partitionnements plats (*flat*) caractérisés par le fait qu'aucune relation précise n'est déterminée entre les différents regroupements générés (figure 2.7 (a)). Les méthodes que l'on retrouve dans cette catégorie sont très souvent de nature itérative. En effet, elles procèdent d'abord en déterminant un nombre fini de regroupements, puis en raffinant chacun des regroupements effectués (Grabmeier et Rudolph, 2002).



(a) (b)

Figure 2.7. Représentations d'un regroupement plat (a) et d'un regroupement hiérarchique (b).

D'autre part, on retrouve aussi des techniques de classification permettant d'effectuer des regroupements hiérarchiques où chaque nœud représente une sous-classe d'un nœud de niveau supérieur. Dans cette perspective, les feuilles représentent les différents objets (documents) classés (figure 2.7 (b)). Les méthodes les plus fréquemment employées pour la classification hiérarchique peuvent être regroupées en deux sous-catégories. La première englobe les méthodes dites « agglomératives » (*bottom-up*). Ces dernières procèdent en identifiant d'abord tous les éléments à classer, puis en effectuant successivement plusieurs regroupements. La seconde englobe les méthodes dites « divisives » (*top-down*). Ces dernières procèdent de manière inverse, c'est-à-dire en identifiant d'abord un seul regroupement puis en divisant ou fractionnant le regroupement initial.

Comme nous l'avons mentionné précédemment, plusieurs méthodes de classification automatique ont été explorées et appliquées avec succès au traitement automatique des documents. Parmi les méthodes les plus fréquemment citées, on retrouve la méthode des k-moyens, les diverses méthodes neuronales (les cartes auto-organisatrices de Kohonen (Kohonen, 2001), le réseau neuronal ART1 (Grossberg et Carpenter, 1987) et ses variantes

ART2 (Grossberg, Carpenter et Rosen, 1991), Fuzzy ART Map (Grossberg *et al.* 1991), etc.), la techniques des plus proches voisins (*nearest neighbor clustering*) (Jain *et al.* 1999; Yang, 1999; Yang et Lui, 1999), les *Support Vector Machines* (SVM) (Joachims, 2002), etc. Dans notre projet, nous avons utilisé le classifieur neuronal ART1.

2.2.3.2.2. Le modèle ART1

Le modèle ART1 est un algorithme de classification (*clustering*) ou de regroupment de type adaptatif, auto-associatif et non supervisé. Il fut développé par S. Grossberg (Grossberg et Carpenter, 1987) en tant que théorie du traitement de l'information. Ce modèle cherche à construire des classes d'information. Il se veut un système d'apprentissage autorégulé. Le principal avantage du modèle ART1 réside dans sa capacité à traiter les intrants de manière dynamique. À cet égard, les implémentations de l'algorithme ART1 possèdent des capacités classificatoires qui peuvent construire par étapes des classes, mais qui, par ailleurs, peuvent aussi s'adapter à un corpus lui-même changeant. Autrement dit, il s'agit d'un réseau de neurones doublement dynamique. Par ailleurs, plusieurs études récentes ont démontré son efficacité à des fins de classification des documents (Massey, 2003a, 2003b).

Ce type d'algorithme possède la particularité de prendre comme intrants un ensemble de vecteurs (c'est-à-dire un ensemble de valeurs caractérisant un objet ou une entité) et, suite à une phase de traitement, de créer des ensembles, des regroupements de vecteurs. Le regroupement s'effectue sur la base de certains critères de similarité. Comme nous le verrons, ces regroupements de vecteurs peuvent facilement faire l'objet d'un étiquetage sémantique ou thématique en fonction de traits caractéristiques partagés par l'ensemble des vecteurs préalablement regroupés.

Très souvent, la représentation interne des regroupements se fait sur la base de vecteurs prototypes. Ces derniers ne sont que des vecteurs dont le rôle consiste à identifier et indiquer la similarité entre les vecteurs intrants regroupés au sein des *clusters*.

2.2.3.2.2.1. Le dilemme entre stabilité et plasticité

Traditionnellement, pour des tâches de classification, plusieurs chercheurs ont favorisé l'utilisation de réseaux neuronaux à rétropropagation (*backpropagation*). Toutefois, les recherches dans ce domaine ont rapidement démontré qu'une telle méthode, tout en effectuant correctement certaines tâches de classification, ne permet pas d'intégrer de nouveaux vecteurs au processus de classification déjà entrepris. Lorsque la classification est effectuée sur un ensemble dynamique d'intrants, il est nécessaire, selon cette approche, de reprendre l'ensemble des calculs déjà effectués pour intégrer les modifications subies par les vecteurs intrants. Et cette opération, lorsque confrontée à d'importants corpus, est « computationnellement » des plus coûteuses. Bref, les algorithmes de rétropropagation sont très efficaces sur des corpus fermés, non évolutifs, mais se heurtent rapidement à un important problème de plasticité.

Afin de contrer cette limite des réseaux à rétropropagation, il est possible de concevoir un réseau de neurones distinct dont la particularité est de ré-effectuer la phase d'entraînement sur les nouveaux vecteurs intrants. Cette caractéristique permet au réseau de s'adapter aux différents changements de son environnement. Par opposition au réseau précédent, ce second réseau possède la capacité de s'adapter à son environnement en traitant de manière dynamique les intrants différents, mais se ressemblant. Nous dirons alors de ce réseau qu'il est « plastique ».

Toutefois, la majorité des réseaux plastiques ne peuvent conserver dans le temps leur apprentissage. Ainsi, la qualité des résultats obtenus par ces réseaux sensibles à l'ajout de nouveaux vecteurs intrants diminuera rapidement dans le temps, au fur et à mesure de l'ajout des nouveaux intrants. En ce sens, cet algorithme ne saurait satisfaire le critère de stabilité, selon lequel le système doit conserver dans le temps les structures reconnues (connaissances acquises) malgré la différence des stimuli intrants.

C'est ce conflit entre la plasticité et la stabilité d'un réseau de neurones qui sous-tend l'ensemble des travaux de Grossberg sur la théorie de la résonance adaptative (*Adaptative Resonance Theory, ART*) (Grossberg et Carpenter, 1987). Ce dilemme entre plasticité et stabilité est posé, de manière plus formelle, dans les termes suivants :

a) Comment un système d'apprentissage, tout en étant conçu pour demeurer plastique ou

adaptatif à l'égard d'informations pertinentes, peut-il demeurer stable à l'égard d'informations non pertinentes?

b) Comment un système peut-il varier adéquatement entre des modes de plasticité et de stabilité afin d'atteindre la stabilité sans être rigide et la plasticité sans être chaotique?

c) Comment est-il possible pour un système de conserver son apprentissage tout en continuant d'apprendre de nouvelles choses?

d) Qu'est-ce qui empêche le nouvel apprentissage d'éliminer ou de supprimer les apprentissages antérieurs?

Ce sont ces quatre questions qui sont à l'origine des travaux de Grossberg, car, jusqu'au milieu des années quatre-vingt, la majorité des algorithmes de classification sont soit stables, mais incapables de générer de nouveaux *clusters* de vecteurs; soit plastiques (donc sensibles aux nouveaux intrants), mais instables dans leur apprentissage.

Un bon système neuronal doit constamment passer d'un mode plastique à un mode stable et *vice versa*. Il doit être en mesure de conserver l'information antérieure, mais en même temps de tenir compte de la nouveauté. Ceci signifie concrètement que le système cherche constamment à s'adapter aux nouveaux intrants tout en conservant les classes antérieures. Il doit donc stabiliser les classes qu'il découvre, mais aussi les changer, si cela est nécessaire, en regard de la réalité nouvelle qui se présente à lui.

Une première solution possible à ce dilemme entre plasticité et stabilité consiste à ce que le réseau à rétropropagation ré-effectue à chaque nouvel intrant le calcul de la phase d'entraînement. Toutefois cette procédure, dont les résultats s'avèrent peu concluants, est, de surcroît, impraticable dans les faits. Au problème initial s'ajoute une considération supplémentaire dont il importe de tenir compte. L'algorithme de classification doit, tout en permettant de solutionner le dilemme de la plasticité et de la stabilité, être de nature incrémentale, c'est-à-dire qu'il devra être sensible à son environnement et accepter de nouveaux intrants tout en reposant sur un apprentissage continu. Pour sa part, le modèle ART1 cherche à contrôler la qualité des intrants et donc à en arriver à une meilleure classification ou auto-organisation.

Combien y aura-t-il de classes produites? Cela est contrôlé par un paramètre de vigilance. Ce paramètre donne une certaine stabilité à la classification. Les dernières classes produites ne détruisent pas les classes antérieures. Il est important de noter que, dans ce modèle, la

classification produite semble malheureusement influencée par l'ordre dans lequel se fait l'apprentissage.

Ce modèle a été modifié et rendu plus efficace dans ART2 (Grossberg *et al.*, 1991). Dans cette version plus complexe, on fait subir à l'échantillon un filtrage qui normalise et élimine le bruit. De plus, les prototypes formés y sont légèrement modifiés et mis à jour en regard de certains paramètres. Cela ne modifie pas en profondeur le modèle, mais accélère le traitement.

Mais, de manière générale, tous les modèles de la famille ART normalisent les intrants, réduisent le bruit et stabilisent les patrons dans le temps. La normalisation de la fonction de transfert consiste à établir un même seuil qui est imposé à tous les neurones. Cette normalisation permet une certaine stabilisation des intrants, surtout lorsque leur impact fluctue de stimulus en stimulus. La réduction du bruit consiste en la définition de paramètres d'intensité avec lesquels un intrant peut agir. Finalement, la stabilisation dans le temps est la possibilité pour le système de considérer un intrant en regard des intrants antérieurs. C'est en ce sens qu'un tel système ne perd pas l'information acquise antérieurement.

Bien qu'il existe plusieurs versions du modèle ART, nous nous attardons ici uniquement sur la première version de celui-ci. Cette version se distingue des autres (ART2, Fuzzy Art) en partie par le fait qu'elle ne traite que des vecteurs intrants de nature binaire et ce de manière non supervisée, c'est-à-dire que l'algorithme établit lui-même les regroupements de vecteurs sans aucune intervention de l'extérieur.

2.2.3.2.2.2. L'architecture

L'idée principale du modèle ART1 est celle d'un système d'interaction entre deux niveaux qui entrent en phase de résonance (figure 2.8).

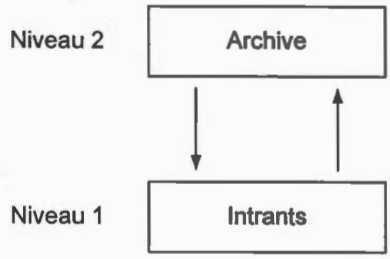


Figure 2.8. L'interaction entre les intrants et le niveau d'archive dans le modèle ART1.

Cette figure représente la relation existant entre les deux niveaux du système. Dans un premier temps, le système reçoit au premier niveau N_1 des stimuli intrants représentant sous forme de vecteurs binaires le premier élément à classer. Par la suite, ces stimuli sont modifiés selon une distribution et un poids particuliers et envoyés au niveau d'archivage N_2.

Au second niveau, le stimulus reçu est donc légèrement différent du stimulus d'origine. Au niveau N_2, le patron est archivé et servira de gabarit (prototype) auquel les intrants suivants seront comparés. À ce moment, ce prototype est retourné (selon les mêmes règles de sommation) au premier niveau.

En termes plus généraux, le patron archivé au niveau N_2 sert d'hypothèse avec laquelle les futurs intrants seront comparés. Ainsi, la comparaison s'effectue entre le patron archivé et les intrants suivants. Dans le cas où le nouvel intrant se distingue radicalement du patron initial (selon un critère ou paramètre de vigilance ρ déterminé par l'utilisateur du système), un nouveau patron sera à son tour créé et servira éventuellement de gabarit aux autres intrants auxquels le système sera possiblement confronté. Dans le cas où le nouvel intrant se présente comme étant relativement comparable au patron initial, il est regroupé (selon des paramètres) avec ce même patron. C'est dans cette perspective qu'il importe de concevoir le phénomène de résonance. Il s'agit de la correspondance entre les patrons prototypes et les patrons intrants. Au fur et à mesure que se poursuit l'apprentissage, une consolidation émerge de cette résonance. L'adaptation se produit dans la modification constante des interconnexions entre les deux niveaux.

2.2.3.2.2.3. L'algorithme

Jusqu'à présent, nous nous sommes limité à présenter de manière générale quelques fondements et principes généraux du modèle ART1. Afin de bien saisir le procédé par lequel ce modèle effectue la classification, nous nous attarderons à présent sur l'algorithme de classification, ainsi que sur les principaux mécanismes de fonctionnement de ce système.

L'algorithme ART1 peut être décrit de manière générale à l'aide de cinq étapes. Dans un premier temps, on retrouve l'étape d'initialisation du système (1). Cette étape est composée de deux processus. Le premier concerne la valeur du paramètre de vigilance. Est donc défini à cette étape le paramètre de vigilance ρ, lequel servira à assurer la stabilité de la classification et agira en tant que critère pour établir le nombre de regroupements d'information. Ce paramètre de vigilance est représenté par une valeur comprise entre zéro (0) et un (1). Ce critère définit la taille des classes à obtenir. Plus la valeur sera proche de zéro (0), plus les classes obtenues seront volumineuses. À l'inverse, une valeur du paramètre de vigilance près de un (1) résultera en une classification plus fine et en une multiplication du nombre de classes obtenues. Dans le modèle ART1, la valeur accordée au paramètre de vigilance provient de l'extérieur du système (donc de l'utilisateur) et est déterminée de manière empirique (c'est-à-dire par essai et erreur).

Le second processus composant cette première phase consiste à entraîner le système afin de déterminer l'ensemble des vecteurs prototypes auxquels seront comparés les intrants (vecteurs) suivants.

La deuxième étape (2) consiste à introduire un nouveau vecteur à classer. Le système initialise alors l'entrée du vecteur suivant et le compare à l'ensemble des vecteurs prototypes candidats. Par la suite, le système identifie le vecteur prototype le plus proche du vecteur intrant (3), après quoi il calcule la distance entre le vecteur prototype sélectionné et le vecteur intrant (4). Finalement, la dernière étape (5) consiste à a) insérer le vecteur intrant dans la classe décrite par le vecteur prototype sélectionné, ajuster le vecteur prototype (dans le cas où le vecteur intrant est suffisamment près du vecteur prototype (en fonction du seuil de vigilance ρ)) et à répéter l'opération à partir de l'étape 2 ou b) ajuster les vecteurs prototypes et reprendre la procédure (3) dans le cas où le vecteur intrant est trop éloigné. De manière schématique, nous obtenons donc la procédure suivante (figure 2.9) :

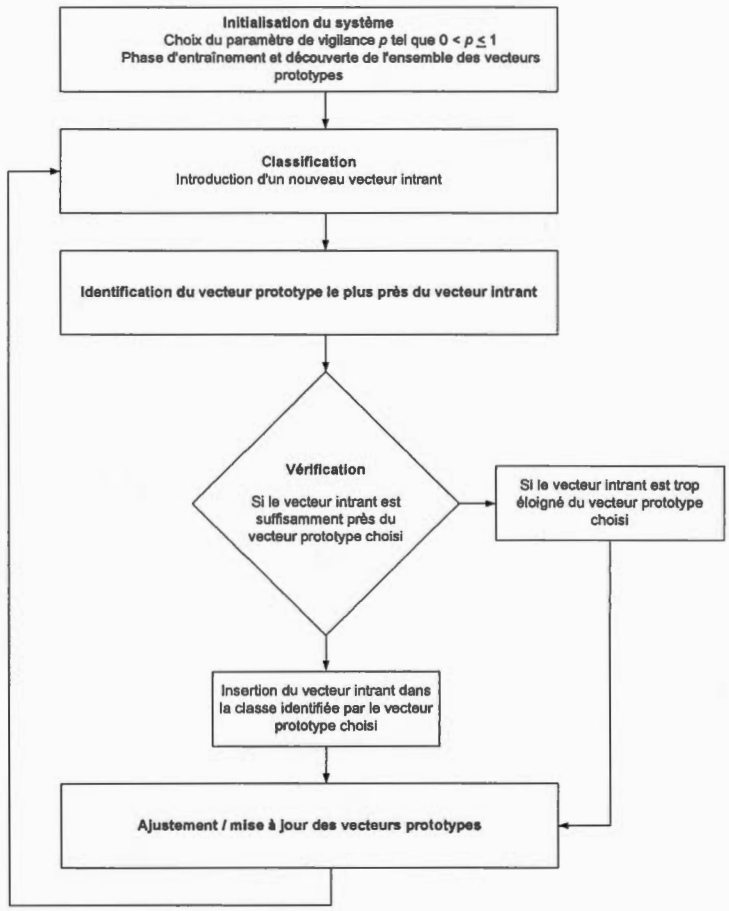


Figure 2.9. Schéma de l'algorithme du modèle ART1.

Les résultats obtenus suite à ce traitement peuvent être représentés dans un espace à N dimensions de la manière suivante (figure 2.10) :

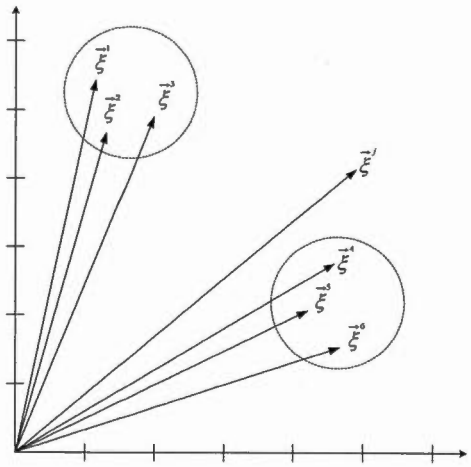


Figure 2.10. Représentation dans un espace à deux dimensions de la classification effectuée par l'algorithme ART1.

2.2.3.2.2.4. Réseaux neuronaux et apprentissage

De manière formelle, l'apprentissage du système ART1 s'effectue au moyen d'un réseau multicouches composé d'une couche d'entrée (qui est aussi une couche de sortie) et d'une couche cachée (figure 2.11). Et cet apprentissage consiste autant dans la détermination des poids des neurones que dans la valeur du seuil de vigilance ρ.

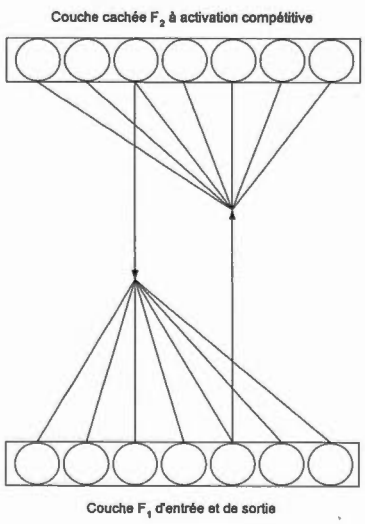


Figure 2.11. Les deux couches du système ART1.

Pour cette phase d'apprentissage, plusieurs processus entrent en jeu (cette section est adaptée de Touzet, 1992):

1. La première phase consiste à initialiser des poids aléatoirement entre zéro (0) et un (1) et à choisir le paramètre de vigilance ρ, tel que $0 < \rho \leq 1$.
2. Par la suite, le système présente un vecteur d'entrée $\vec{\xi}^{e}$ appartenant à la base d'apprentissage.
3. À ce vecteur d'entrée $\vec{\xi}^{el}$ est associé un neurone gagnant sur la couche cachée N_j.
4. Suite à ce calcul est généré un vecteur binaire de sortie $\vec{\xi}^{sj}$ issu de ce seul neurone gagnant $\vec{\xi}^{nj}$.
5. Après quoi, le système tente d'unifier le vecteur de sortie $\vec{\xi}^{sj}$ et le vecteur d'entrée $\vec{\xi}^{el}$. Si $|\vec{\xi}^{sj}| / |\vec{\xi}^{el}| \geq \rho$ (où $|\vec{\xi}^{sj}|$ est la norme du vecteur $\vec{\xi}^{sj}$, laquelle est égale au nombre de composantes un (1)), alors l'unification est réalisée. Il faut alors ajuster les poids (passer à l'étape 7).
6. Sinon, c'est-à-dire si $|\vec{\xi}^{sj}| / |\vec{\xi}^{el}| \leq \rho$, le neurone gagnant $\vec{\xi}^{nj}$ est inhibé. S'il y a encore des neurones non inhibés sur la couche cachée, alors il faut retourner à l'étape 3. Sinon, un nouveau neurone caché (représentant le vecteur prototype de la nouvelle classe) est créé et initialisé comme représentant de la classe correspondant à la forme d'entrée $\vec{\xi}^{el}$ en utilisant la modification des poids décrite à l'étape 7.
7. L'avant-dernière étape consiste à modifier les poids :
   7.1. Pour la couche des poids montants, où $_h$ est un neurone de la couche d'entrée et $_j$ un neurone gagnant de la couche cachée, la modification s'effectue de la manière suivante :
      - Si le neurone $_h$ est actif (s'il possède la valeur un (1)) : $W_{jh} = 1 / |\vec{\xi}^{sj}|$
      - Si le neurone $_h$ est inactif (s'il possède la valeur zéro (0)) : $W_{jh} = 0$

   7.2. Pour la couche des poids descendants, où $_j$ est un neurone de la couche cachée et $_k$ un neurone gagnant de la couche d'entrée, la modification s'effectue alors comme suit :

- Si le neurone est actif : $W_{kj} = 1$
- Si le neurone est inactif et possède donc la valeur zéro (0) : $W_{kj} = 0$

7.3. Retour à l'étape 2.

8. Quand le passage de tous les exemples de la base d'apprentissage n'occasionne plus aucun ajout de neurone, il peut être utile de mesurer les performances en contrôlant le nombre de classes et la qualité des classes construites. Si le nombre de classes est trop faible (c'est-à-dire si la classification n'est pas suffisamment fine), on retourne à l'étape 1 avec une augmentation de la valeur de ρ. Si le nombre de classes est trop élevé (c'est-à-dire si la classification s'est avérée trop fine), on retourne à l'étape 1 en diminuant la valeur de ρ. Il est à noter que, pour ce paramètre de vigilance, peu d'indications précises sont disponibles. Il faut donc nécessairement déterminer la valeur optimale du paramètre de vigilance de manière empirique. Et cette opération est des plus complexes, car la valeur optimale du paramètre de vigilance varie selon plusieurs facteurs, telle la taille et l'ordre de présentation des vecteurs, la nature des objets représentés, etc.

La figure 2.12 présente comment un vecteur intrant particulier est soumis au système (il en va de même pour l'ensemble des vecteurs à classer lors de la phase d'apprentissage). Dans un premier temps, après un calcul et une compétition entre les neurones de la couche cachée, un seul neurone gagnant de la couche F_2 est choisi. Il s'agit dans cet exemple du neurone j. Il s'agit donc, selon le système, du neurone le plus représentatif du vecteur binaire d'entrée $\vec{\xi}^{e}$ (figure 8 (a)). À son tour, le neurone gagnant j génère sur la couche de sortie un vecteur binaire prototype $\vec{\xi}^{s}$. Ce dernier est par la suite comparé au vecteur d'entrée $\vec{\xi}^{e}$ (figure 8 (b)). Après ces étapes, deux scénarios sont envisageables. Si la différence entre les vecteurs $\vec{\xi}^{s}$ et $\vec{\xi}^{e}$ est inférieure à un seuil déterminé, le neurone gagnant est alors considéré comme représentatif de la classe du vecteur d'entrée $\vec{\xi}^{e}$. Suit alors une modification des poids de connexion du neurone gagnant, modification qui a pour effet de consolider l'apprentissage en renforçant les liens d'activation entre j et $\vec{\xi}^{e}$. Si, à l'inverse, la différence entre les vecteurs $\vec{\xi}^{s}$ et $\vec{\xi}^{e}$ est supérieure à un seuil déterminé, il y a reprise du calcul et de la compétition entre tous les neurones de la couche cachée, moins le neurone gagnant de l'étape

précédente (figure 8 (c)). Ce processus se poursuivra jusqu'à ce que le premier scénario soit réalisé (figure 8 (d)). Lorsque tous les neurones de la couche cachée sont passés en revue sans qu'il n'y ait aucune adéquation ou association avec le vecteur d'entrée ξ^{e}, un nouveau neurone caché est ajouté et initialisé comme représentant de la classe du vecteur d'entrée ξ^{e}.

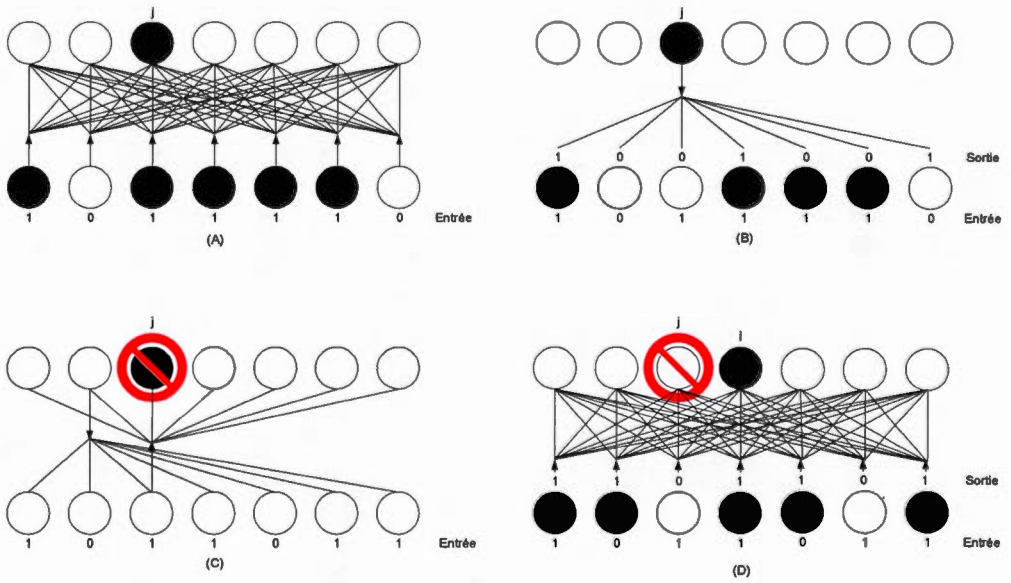


Figure 2.12. Le fonctionnement de ART1.

2.2.3.2.2.5. Les limites du réseau ART1

Malgré ses nombreux avantages, le réseau ART possède certains inconvénients. Le premier de ceux-ci est que la classification produite par ce réseau semble grandement influencée par l'ordre dans lequel se fait l'apprentissage. De plus, ART1 est incapable de traiter des vecteurs intrants composés de valeurs pondérées. En effet, le modèle ART, dans sa version initiale, n'accepte comme intrants que des vecteurs binaires. Cette caractéristique de l'algorithme ART1, lorsqu'il est appliqué au traitement de données textuelles, se manifeste par la présence de vecteurs creux, ce qui alourdit significativement le traitement des données et le travail d'interprétation des résultats obtenus. D'ailleurs, d'un point de vue cognitif, cette caractéristique soulève un problème théorique important. En effet, lorsque nous nous

interrogeons sur les traits caractéristiques d'un véritable système d'apprentissage, on constate qu'en plus d'avoir à satisfaire plusieurs critères (capacité d'organisation, d'apprentissage continu, etc.), il ne doit théoriquement pas poser de restrictions quant à la forme que peuvent prendre les signaux intrants. Et c'est justement à cet égard que le modèle ART1 pose problème : en raison de son architecture, il n'accepte que des données binaires.

D'autre part, il serait justifié d'attendre d'un système d'auto-apprentissage et d'auto-organisation des données qu'il effectue un traitement sans intervention provenant de l'extérieur du système. À cet égard, le système ART1 semble à nouveau poser problème, car la définition du paramètre de vigilance ρ doit être effectuée par l'utilisateur du système. C'est lui qui doit manuellement entrer la valeur qu'il désire attribuer à ce paramètre.

Finalement, une dernière remarque mérite d'être faite. Elle concerne une caractéristique de la classification obtenue par le système ART1. Comme nous l'avons souligné lors de la présentation du modèle, ART1 effectue une classification exclusive. Nous entendons par ce terme une classification où un même vecteur intrant ne peut appartenir qu'à une seule classe.

Dans le cadre de notre projet, la classification des segments de documents a été effectuée en employant une implémentation de l'algorithme ART1 réalisée par notre collègue Sébastien Hélie.

Les résultats obtenus par le processus de classification prennent la forme de classes de segments (figure 2.13), desquelles il est possible d'extraire le lexique.

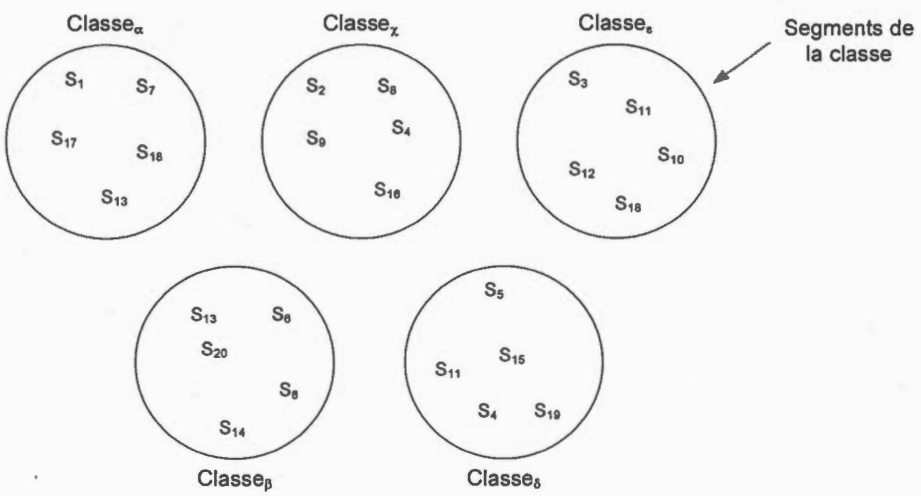


Figure 2.13. Le résultat de la classification : des classes de segments.

2.2.3.3. L'extraction automatique des termes thématiques

Comme nous l'avons évoqué précédemment, dans bon nombre d'applications d'AGIT et de LATAO, l'opération de catégorisation est effectuée en utilisant un plan de classification ou une taxinomie de catégories prédéfinies (Jackson et Moulinier, 2002; Manning et Schütze, 1999; Sebastiani, 1999). Ce fait est d'ailleurs clairement évoqué dans plusieurs travaux sur l'analyse des données textuelles.

> L'opération centrale [de la catégorisation] tient dans l'élaboration d'une grille de catégories. Il s'agit en effet d'enregistrer tous les éléments du corpus pertinent afin de les classer par thèmes ou catégories thématiques, souvent en vue d'établir des pourcentages et de procéder à des comparaisons significatives entre les différents documents-support. (Robert et Bouillaguet, 1997, p. 27).

Ceci est le cas de l'application NEFCLASS-J. En effet, cette dernière exige que l'ensemble des données qui lui sont soumises durant le processus d'apprentissage soit préalablement catégorisé selon une taxinomie prédéfinie et adaptée au corpus à traiter.

Cependant, le développement des taxinomies, bien que de plus en plus de projets de recherche tentent de l'assister à l'aide d'applications informatiques, s'avère très coûteux (en temps et en ressources).

> Catégoriser suppose évidemment que l'on ait compris; la compréhension sera d'autant plus fine que le chercheur sera familiarisé, au début de son étude, sinon déjà avec le corpus lui-même, du moins avec l'« univers mental » dont fait partie le corpus. Définir les catégories revient à expliciter la compréhension intérieure que l'on a du contenu sémantique global de la « base du texte » ou message-source. (Robert et Bouillaguet, 1997, pp. 27-28).

De plus, comme l'ont souligné entre autres Louwerse et Van Peer (2002, p. 4), les index thématiques (i.e. les catégories thématiques prédéfinies) posent plusieurs problèmes : « *The index was conceived to be a practical reference, but trying to classify tales in the [...] index proved problematic.* »

Un des objectifs de notre projet est de proposer une solution efficace mais peu coûteuse à ce problème. Nous avons expérimenté une technique de catégorisation thématique fondée sur l'extraction automatique de termes thématiques à partir des documents classifiés. Afin de

dépasser les limites de la catégorisation automatique effectuée à partir d'ensembles de catégories thématiques prédéfinies, nous avons exploité certaines mesures statistiques permettant de faire émerger les catégories thématiques à partir des documents regroupés lors du processus de classification. Notre méthode consiste à appliquer certains critères statistiques utilisés dans les domaines du repérage de l'information (Salton, 1989, Baeza-Yates et Ribeiro, 1999) à chacune des classes lexicales différenciées (obtenues par le processus de classification) afin d'identifier au sein de chacune de ces sous-classes les quelques termes les plus significatifs pouvant (suite à une évaluation) servir d'étiquette thématique pour la découverte des principaux thèmes d'un corpus. De très récents travaux dans le domaine du forage de textes ont d'ailleurs évoqué la pertinence d'explorer différentes techniques visant à extraire automatiquement des informations représentatives à partir de documents non-catégorisés.

> *Unlabeled documents can be useful for prediction even when labels will never be assigned. [...]. A general idea is to find features in unlabeled data. We may discover patterns from unlabeled data and use such patterns as features input to the learning algorithm. If such patterns exist, then unlabeled data can help.* (Weiss, *et al.*, 2005, p. 202)

En outre, cette idée a été évoquée dans un contexte de classification des documents étroitement comparable à celui décrit dans ce travail :

> *We should not lose sight of what we are clustering. Documents are composed of words, and the distribution of words is the basis of document clustering. Surely, we can use these words to express the meaning of the result of the clustering. Documents are clustered by similarity of words, so if we characterize a cluster by the right words, we should be able to give the meaning or rationale for the reasonableness of the cluster.* (Weiss, *et al.*, 2005, p. 121)

Dans ce projet, nous nous sommes limité à appliquer deux mesures computationnellement peu coûteuses : 1) la fréquence relative des termes dans chacun des segments et 2) la mesure tf · idf (« *term frequency · inverse document frequency* ») (Salton, 1989). Le principe de cette seconde mesure peut être formulé de la manière suivante : un terme sera d'autant meilleur pour représenter le contenu d'une classe s'il est à la fois fréquent dans cette classe et rare dans l'ensemble des classes à analyser. La fréquence inverse du document (formule 1) vient

donc modérer ou accentuer l'importance de la fréquence de chaque terme. Ainsi, ce calcul est utilisé, dans le cadre de notre analyse, afin d'extraire les termes les plus représentatifs des classes obtenues. Les termes retenus suite à ce calcul sont alors attribués comme « étiquette thématique » à leur(s) classe(s) respective(s).

$$w_{ik} = \frac{tf_{ik} \log(N / n_k)}{\sqrt{\sum_{k=1}^{t} (tf_{ik})^2 [\log(N / n_k)]^2}} \qquad (1)$$

Dans cette formule, T_k correspond au terme k dans le document D_i; tf_{ik} correspond à la fréquence du terme T_k dans le document D_i; idf_k correspond à la fréquence inverse du terme T_k dans le corpus C; N correspond au nombre de documents dans le corpus C; n_k correspond au nombre de documents dans le corpus C contenant le terme T_k et idf_k correspond au $\log(n_k/N)$.

2.2.4. La découverte et l'analyse thématique des documents

La découverte des thèmes et l'analyse thématique reposent sur les résultats des processus de classification et de catégorisation. L'analyse thématique, par opposition à l'identification des thèmes d'un corpus, est définie, dans le cadre de notre projet, comme un processus de découverte et de parcours des différents thèmes présents dans un corpus textuel. Cette démarche est potentiellement multiple et fort complexe. Elle repose en dernière instance sur plusieurs choix, tant théoriques que pratiques. Mais, de manière générale, la navigation thématique consiste en un parcours caractérisé par un compromis entre, d'une part, les attentes du lecteur et, d'autre part, les indices sémiotiques présents dans le texte. Comme le souligne Bremond (1985, p. 420) :

> Par quoi suis-je orienté dans la série de mes choix? On peut répondre : par le désir d'isoler la ou les bonnes formes du thème. Mais qu'est-ce qu'une bonne forme? [...] la bonne forme, c'est celle qui procure la satisfaction la plus grande à mon attente de lecteur [...].

Pour d'autres (Prince, 1985, p. 430), cette attente du lecteur prendra le nom de « réalité extra-textuelle ». D'ailleurs plusieurs théoriciens ont noté l'importance, dans l'activité de

thématisation et de découverte des contenus thématiques, de cette composante essentiellement subjective. Martin (1995) opte pour une position analogue à celle de Prince, lorsqu'il écrit que « la nature de ce que l'on pourrait appeler plus généralement l'étude thématique des textes est d'abord fonction de l'objectif visé » (p. 18).

Dans le cadre de notre recherche, la composante subjective se manifeste dans l'intérêt du chercheur envers certains thèmes qu'il privilégie et dans les objectifs qu'il désire atteindre lors de son analyse. Ainsi, le chercheur peut, par exemple, choisir d'explorer un ou plusieurs thèmes précis du corpus, et ce dans le but d'en démontrer l'organisation, la structure, etc. Peut-être voudra-t-il explorer l'ensemble des thèmes d'un corpus afin d'orienter ou de cibler les passages qu'il voudra analyser plus en détail par la suite.

> Thématiser un texte dépend donc non seulement du « texte même » mais aussi (et peut-être davantage) du thématiseur, du cadre adopté, des unités choisies, des opérations accomplies pour les harmoniser, des résumés et paraphrases effectués. (Prince, 1985, p. 432)

Mais, d'autre part, une seconde contrainte, plus objective cette fois, entre aussi en compte dans le cadre de la tâche d'analyse et de découverte. Cette contrainte repose sur le texte à analyser[9]. Cette composante intra-textuelle limite nécessairement la liberté de l'interprète, car elle guide inévitablement l'ensemble des analyses. En effet, malgré les intérêts et les raisons qui mènent le chercheur vers la découverte d'un thème particulier plutôt que d'un autre, le chercheur ne crée pas entièrement les thèmes dans le corpus qu'il analyse. C'est le texte qui expose, à l'aide des différents porteurs sémiotiques qu'il comporte, les thèmes sur lesquels le chercheur posera éventuellement son analyse.

De fait, à partir des démarches méthodologiques dont nous avons précédemment décrit en détail les étapes, le processus d'analyse thématique se déroule ainsi. Dans un premier temps, le chercheur, ayant obtenu les résultats de la catégorisation ou de la classification des multiples segments de textes issus de son corpus de départ, procédera à l'analyse du lexique de chacune des classes catégorisées.

---

[9] Comme le note Bremond : « À moins, bien sûr, que le texte, par un jeu d'indices sémiotiques, n'oriente ma conceptualisation du thème [...] » (Bremond, 1985, p. 420).

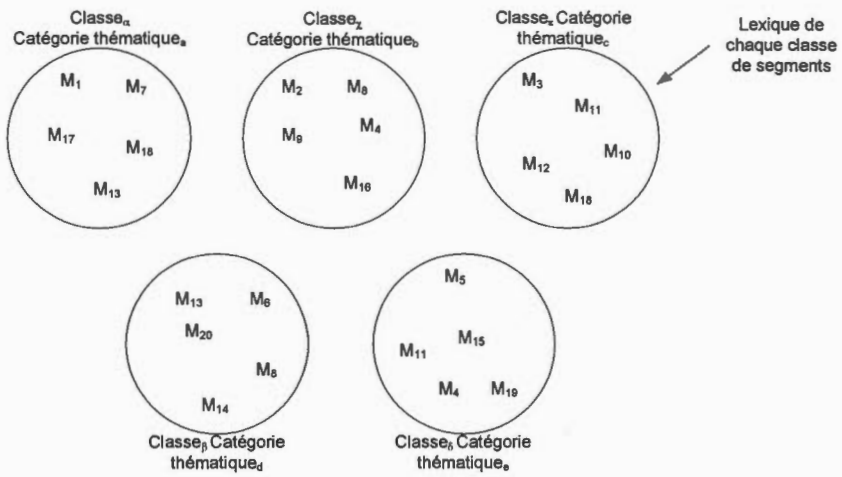


Figure 2.14. Représentation graphique du lexique de chaque classe catégorisée.

Dans cette figure (2.14), chaque classe est constituée d'une liste de lexèmes (M). Plusieurs cheminements sont alors possibles pour une analyse thématique. Certaines unités lexicales présentes dans une classe particulière peuvent se retrouver aussi dans une autre classe, indiquant par là qu'elles opèrent dans un autre contexte. Ainsi, dans une classe de départ, le lecteur peut partir d'un terme choisi pour son intérêt thématique et naviguer dans une autre classe où le même terme se retrouve, mais cette fois dans un nouveau contexte. Ce contexte, à son tour, est constitué de lexèmes nouveaux qui peuvent servir de départ pour aller vers d'autres classes. Et ce processus recommence indéfiniment jusqu'à la clôture ou la saturation du parcours. Ainsi, au terme de son parcours, le lecteur aura exploré l'ensemble de son corpus textuel, de segment en segment (regroupés sous forme de classes), mais sans nécessairement savoir au préalable vers quel but. Le parcours est heuristique et s'adapte aux résultats obtenus. De nombreux chemins sont possibles, ouvrant l'analyse thématique vers de multiples horizons (voir figure 2.15).

> À l'examen des terrains de chaque thème s'ajoute celui de leur(s) interaction(s), et les uns et les autres diffèrent selon les places et rôles respectifs des thèmes considérés. (Martin, 1995, p. 22)

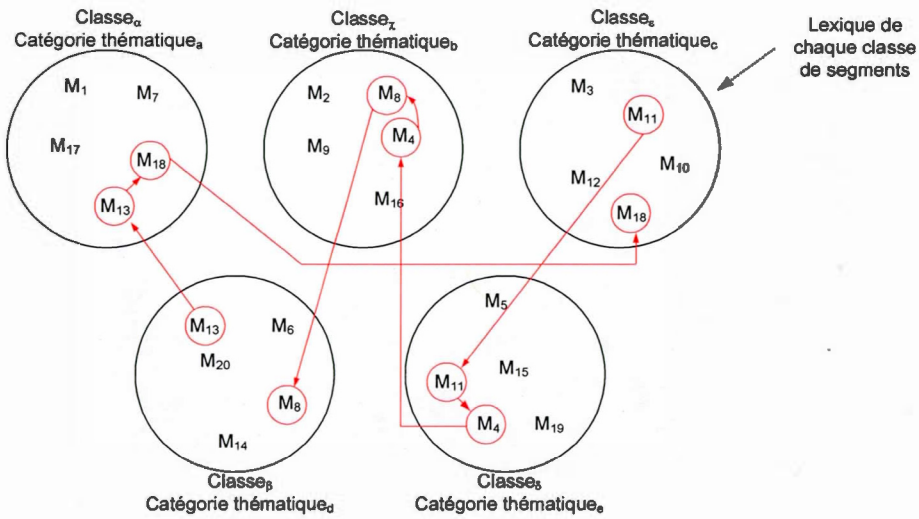


Figure 2.15. La navigation thématique et la découverte des thèmes d'un corpus.

Le potentiel applicatif d'une telle démarche repose en partie sur la manière dont les résultats sont présentés à l'utilisateur. Les résultats générés sous forme de listes ne sont manifestement pas des plus conviviaux. Par conséquent, la représentation graphique des résultats dans les domaines du traitement informatisé des données textuelles constitue actuellement le lieu de nombreuses recherches. Certaines d'entre elles ont récemment proposé des applications permettant d'encoder manuellement (en utilisant très souvent la norme XTM 1.0 (*XML Topic Maps*)) et de visualiser les principaux thèmes d'un corpus de données textuelles.

Dans notre projet, nous avons retenu l'application de représentation graphique UCINET (Borgatti, Everett et Freeman, 2002) (fondée sur la théorie des réseaux sociaux) à des fins de présentation ergonomique des résultats. Avant d'arrêter notre choix sur cette application, nous avons préalablement exploré plusieurs outils de visualisation dont GRAPHVIZ (www.graphviz.org), HYPERGRAPH (hypergraph.sourceforge.net), INFOVIZ (ivtk.sourceforge.net), PREFUSE (prefuse.sourceforge.net) et TULIP (www.tulip-software.org). Plusieurs des outils que nous avons explorés génèrent des représentations complexes, mais peu d'entre eux comportent des fonctionnalités permettant d'automatiser le processus de représentation. Dans ces applications, c'est l'utilisateur qui doit encoder manuellement l'ensemble des données qu'il souhaite visualiser. Nous avons donc retenu l'application

UCINET, car le module de représentation de ce logiciel permet de contourner cette importante limite. De plus, comme plusieurs l'ont démontré (voir entre autres Archambeault, 2002; Popping, 2000), les outils de représentation fondés sur la théorie des réseaux sociaux sont bien adaptés pour représenter les données textuelles traduites dans un modèle vectoriel.

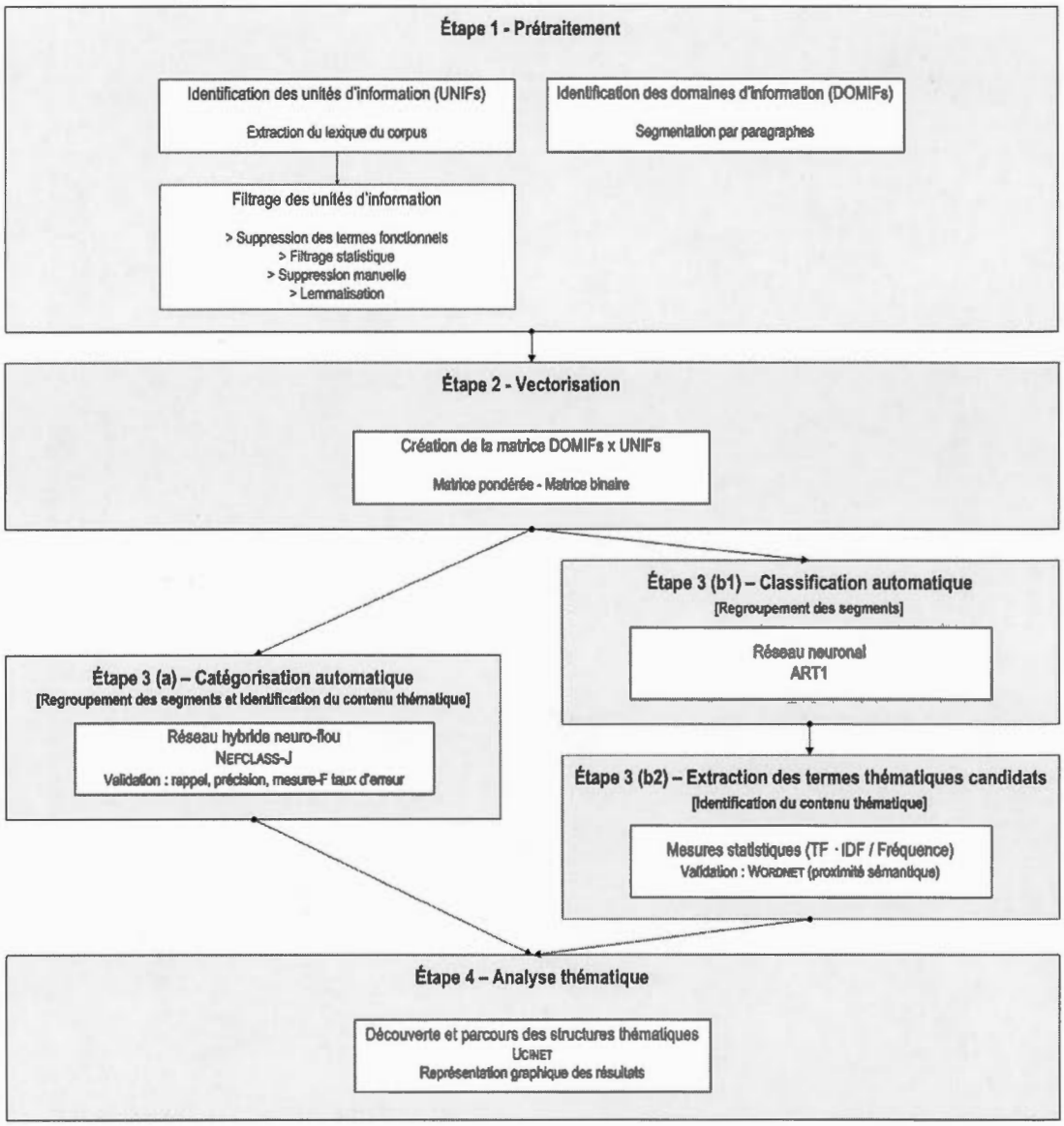


Figure 2.16. Schéma récapitulatif de la démarche proposée.

CHAPITRE 3

EXPÉRIMENTATION ET RÉSULTATS

3.1. Corpus

L'expérimentation et la validation des deux démarches méthodologiques ont été réalisées sur un échantillon du corpus d'articles du journal belge *LE SOIR*. Dans les domaines du repérage de l'information et du traitement automatique du langage, plusieurs corpus de référence ont été développés durant les dernières années. À des fins de repérage d'informations, la collection de corpus la plus fréquemment citée est sans aucun doute celle du projet TREC[1] (Voorhees et Harman, 2005). Les corpus de ce projet sont d'abord caractérisés par la quantité importante de documents qu'ils contiennent (chaque corpus contient entre 500 000 et 1 000 000 de documents). Les corpus se distinguent par les différents traitements qu'ils permettent d'évaluer. En effet, le projet TREC est divisé en plusieurs volets (*tracks*) focalisés sur la validation de tâches spécifiques reliées au domaine du repérage de l'information. Ainsi, on retrouve dans ce projet des corpus visant à évaluer des systèmes permettant d'identifier de nouvelles informations (*novelty track*), de retrouver des réponses à des questions précises (*question answering track*), etc[2]. Tous les corpus TREC sont composés, indépendamment des traitements auxquels ils sont destinés, de trois principales sections : un ensemble de documents (ensemble d'apprentissage), un énoncé des objectifs à atteindre (nommé *topics*) et d'un second ensemble de documents (ensemble de test). Malheureusement, aucun corpus de la collection TREC n'a été conçu afin de valider des prototypes d'applications visant à identifier le contenu thématique et à assister l'analyse

---

[1] trec.nist.gov

[2] Pour une description détaillée des différents volets de l'édition 2004 du projet TREC, nous référons le lecteur aux documents disponibles à l'adresse trec.nist.gov/pubs.html.

thématique des documents. Il nous a donc été impossible d'employer cette ressource dans le cadre de notre projet.

Par ailleurs, dans les domaines de l'identification du contenu thématique et de l'analyse thématique de documents, aucun corpus de référence n'a été développé et ce, malgré le fait que l'assistance informatique à l'identification du contenu thématique constitue présentement un important territoire de recherche. Le seul corpus que nous sommes en mesure de relier indirectement à notre projet de recherche provient du projet *Topic Detection and Tracking (TDT)*. Ce projet, mis sur pied en 1996, vise à développer et à valider différentes technologies « intelligentes » pour la compréhension de nouvelles journalistiques. Le corpus développé pour la cinquième phase de ce projet est multilingue. Il est en effet composé de 72 910 documents rédigés en arabe, 278 109 documents rédigé en anglais et 56 486 documents rédigés en mandarin (le corpus complet est donc composé de 407 505 documents). Le projet vise à développer et à valider certaines technologies destinées à l'identification du contenu thématique de ces documents. Cependant, les objectifs de notre projet se distinguent de ceux poursuivis par le projet TDT par la manière dont sont définis les thèmes des documents. En effet, dans le cadre du TDT, les thèmes sont définis en tant qu'événements ou activités (ensemble d'événements reliés) associés à un moment particulier, à un lieu particulier, ainsi qu'à des conditions et des conséquences précises. En outre, les thèmes doivent être identifiés en fonction d'une taxinomie composée de 13 catégories thématiques très générales (telles que « *Acts of Violence or War* » ou « *Celebrity and Human Interest News* »).

> *A TDT event is defined as a particular thing that happens at a specific time and place, along with all necessary preconditions and unavoidable consequences. A TDT event might be a particular plane crash, or a single meeting, or a particular court hearing. An activity is a connected set of events that have a common focus or purpose, happening at a specific place and time; for instance, a campaign, or an investigation, or a disaster relief effort. For the purposes of TDT, a topic is defined as an event or activity, along with all directly related events and activities.* (TDT, 2004, p. 4)

Le corpus a donc été conçu afin de permettre d'atteindre des objectifs du TDT. Il nous aurait été impossible d'employer ce corpus, car il ne comporte pas d'informations permettant d'évaluer les résultats obtenus par notre méthodologie. En effet, le projet TDT repose sur une conception bien particulière des concepts de « thème » et de « contenu thématique » ne correspondant pas avec celle de notre projet.

> *It is important to highlight the difference between a TDT topic and the notion of topic in normal discourse. While one might normally think of a topic as something broad like "accidents", a TDT topic is limited to a specific accident, like the cable car crash.* (TDT, 2004, p. 5)

Ne disposant pas de corpus de référence adapté aux spécificités de notre projet, nous avons donc décidé de construire notre propre corpus d'expérimentation. Nous avons opté pour un corpus composé d'articles du journal belge *LE SOIR*. Ce choix est motivé par plusieurs facteurs. D'abord, il s'agit d'un corpus d'articles non-fictifs nous permettant, ainsi, de valider nos hypothèses dans un contexte réel. Les données employées n'ont fait l'objet d'aucune manipulation préliminaire qui aurait pu affecter la qualité des résultats obtenus. Par ailleurs, comme notre démarche fait appel à certaines opérations de nature linguistique (ex. : l'opération de lemmatisation), nous avons jugé préférable d'employer un corpus monolingue (le corpus retenu satisfait cette exigence). Finalement, dans la version numérique du journal *LE SOIR,* chaque article est associé à un ensemble de métadonnées pertinentes pour l'évaluation de notre démarche. En effet, chaque article a été manuellement catégorisé par les éditeurs du journal. Ainsi, cela nous a permis d'identifier la véritable catégorie thématique de chaque article. Nous avons utilisé cette information à des fins d'évaluation des résultats obtenus par les processus de catégorisation automatique et d'extraction automatique des termes thématiques candidats.

La collection complète des archives du journal *LE SOIR* est disponible sur 10 CD-ROM couvrant la période 1994 à 2003. Tous les articles disponibles sur le CD-ROM du journal *LE SOIR* sont non-structurés (il s'agit donc de textes bruts) et catégorisés selon une taxinomie propre à ce journal. Cette taxinomie est constituée de trois niveaux hiérarchiques. Le premier niveau est composé de 21 catégories, le deuxième de 336 catégories et le troisième de 910 catégories. Par ailleurs, il importe de mentionner que la catégorisation effectuée par les éditeurs de la version du journal *LE SOIR* est non-exclusive. Ainsi, la majorité des articles sont associés à plusieurs catégories thématiques.

Nous avions d'abord voulu limiter notre analyse à une seule année (la plus récente disponible sur CD-ROM au moment où nous avons entamé les expérimentations, soit l'année 2002), mais nous avons dû restreindre encore davantage notre corpus, principalement en raison de la grande quantité d'informations disponibles pour cette seule année. En effet, la

version numérique 2002 du journal est composée de 28 552 articles, de plus de 20 000 000 de mots et fait plus de 120 Mo en format texte brut. Il aurait été pratiquement impossible d'analyser rigoureusement un corpus d'une telle ampleur dans le cadre de ce projet. Nous avons donc limité notre corpus d'expérimentation à 250 articles regroupés en 10 catégories de taille égale. La seule contrainte qui a guidé la composition du corpus est au niveau de la taille des catégories. Ainsi, afin d'être suffisamment représentées, les catégories devaient comporter un nombre minimum d'articles. Nous avons établi le seuil minimum d'articles que devaient comporter chaque catégorie à 25. Les catégories retenues, choisies de manière aléatoire à partir des 910 catégories de niveau 3, sont les suivantes : « CINÉMA », « ÉLECTRICITÉ », « GARE », « GASTRONOMIE », « INFORMATIQUE », « ISLAM », « MÉDECIN », « PÉDOPHILIE », « TENNIS » et « UNIVERSITÉ ».

Pour constituer notre corpus, nous avons donc effectué 10 requêtes thématiques afin de récupérer 25 articles de chaque catégorie. À titre indicatif, mentionnons que notre corpus ainsi constitué comporte 158 630 occurrences et 18 616 formes dont la fréquence varie entre 1 et 8 277 occurrences. On y retrouve, entre autres, 9 367 hapax (mots dont la fréquence est de 1). Le ratio formes/occurrences est donc de 11,74%; alors que le ratio hapax/formes est de 50,32%.

Après avoir supprimé les termes fonctionnels, les marques de ponctuation et tous les caractères numériques, la taille du corpus a été substantiellement réduite. Ainsi, le corpus, partiellement nettoyé, est composé de 70 407 occurrences et de 17 818 formes dont la fréquence varie entre 1 et 228 occurrences. On y retrouve désormais 9 145 hapax. Le ratio formes/occurrences est augmenté à 25,31%; alors que le ratio hapax/formes est maintenant de 51,32%. La figure suivante représente un article typique de notre corpus.

Comme nous l'avons mentionné précédemment, nous avons réalisé deux principales expérimentations, nous permettant ainsi d'évaluer deux démarches méthodologiques distinctes (la première, de nature prédictive, fondée sur l'opération de catégorisation automatique; la seconde, de nature exploratoire, fondée sur les opérations de classification automatique et d'extraction automatique des termes thématiques candidats).

Actualité sportive
Mardi 31 décembre 2002 N 304
Page 23

Coupe Hopman
Le sourire était de rigueur à Perth
Clijsters et Malisse ont repris

Par AFP; BELGA

Coupe Hopman
Le sourire était de rigueur à Perth
Clijsters et Malisse ont repris

En offrant quelques friandises à Xavier Malisse, Kim Clijsters pensait pouvoir lui donner l'énergie nécessaire pour clôturer le double mixte décisif pour la Belgique face à l'Espagne pour son premier match de poule de la Coupe Hopman, lundi à Perth (Australie). De fait, le sursaut salvateur dans le super jeu décisif a permis aux Belges de décrocher une première victoire. Auparavant, Clijsters s'était défaite de Virginia Ruano alors que Malisse s'était incliné devant Robredo. Le double a offert le succès.

Devant l'Ouzbékistan jeudi et les Etats-Unis (avec Serena Williams et Blake) vendredi, la Belgique tentera de se hisser en finale en décrochant la première place de ce groupe A.

Je lui ai donné un bonbon, et il a eu un peu plus d'énergie, a expliqué Clijsters. Nous étions menés 4-2, je crois que cela a aidé. Ce qu'a confirmé Malisse. J'étais fatigué, le voyage avait été long. Mais les choses ont aussi commencé à aller mieux après s'être détendu et avoir ri un bon coup. Pour sa part, Clijsters est apparue déjà bien affûtée. Je me sens vraiment très bien. La victoire au Masters à Los Angeles m'a donné du cœur à l'ouvrage. J'ai travaillé encore plus dur. (D'après Belga, DPA, AFP.) .

TYPE D'ARTICLE : Informations, comptes-rendus, récits
MOTS-CLÉS :
- THEMATIQUES : **TENNIS**; COMPETITION SPORTIVE; MATCH; COUPE; BELGIQUE; ESPAGNE; RESULTATS
- NOMS DIVERS : COUPE HOPMAN
NODOC: SR_20021231_475

Figure 3.1. Exemple d'article constituant notre corpus (le terme en gras indique la catégorie thématique à partir de laquelle l'article a été récupéré).

Il est à mentionner qu'avant de soumettre notre corpus à tout traitement, nous avons d'abord supprimé manuellement l'ensemble des métadonnées non-pertinentes à l'analyse. Ainsi, nous avons manuellement supprimé les informations telles que le titre de l'article, le nom de l'auteur, la date de parution, la mention de droits d'auteur, etc. Compte tenu de l'objectif poursuivi dans le cadre de cette recherche, les métadonnées supprimées ne comportaient que des informations qui auraient pu biaiser les résultats des opérations de

catégorisation et de classification[3]. Ainsi, dans l'article présenté à la page précédente, nous avons supprimé les passages en gris, pour ne retenir que les passages en noir (figure 3.2).

Actualité sportive
Mardi 31 décembre 2002 N 304
Page 23

Coupe Hopman
Le sourire était de rigueur à Perth
Clijsters et Malisse ont repris

Par AFP; BELGA

Coupe Hopman
Le sourire était de rigueur à Perth
Clijsters et Malisse ont repris

En offrant quelques friandises à Xavier Malisse, Kim Clijsters pensait pouvoir lui donner l'énergie nécessaire pour clôturer le double mixte décisif pour la Belgique face à l'Espagne pour son premier match de poule de la Coupe Hopman, lundi à Perth (Australie). De fait, le sursaut salvateur dans le super jeu décisif a permis aux Belges de décrocher une première victoire. Auparavant, Clijsters s'était défaite de Virginia Ruano alors que Malisse s'était incliné devant Robredo. Le double a offert le succès.

Devant l'Ouzbékistan jeudi et les Etats-Unis (avec Serena Williams et Blake) vendredi, la Belgique tentera de se hisser en finale en décrochant la première place de ce groupe A.

Je lui ai donné un bonbon, et il a eu un peu plus d'énergie, a expliqué Clijsters. Nous étions menés 4-2, je crois que cela a aidé. Ce qu'a confirmé Malisse. J'étais fatigué, le voyage avait été long. Mais les choses ont aussi commencé à aller mieux après s'être détendu et avoir ri un bon coup. Pour sa part, Clijsters est apparue déjà bien affûtée. Je me sens vraiment très bien. La victoire au Masters à Los Angeles m'a donné du cœur à l'ouvrage. J'ai travaillé encore plus dur. (D'après Belga, DPA, AFP.)

TYPE D'ARTICLE : Informations, comptes-rendus, récits
MOTS-CLÉS :
- THEMATIQUES : TENNIS; COMPETITION SPORTIVE; MATCH; COUPE; BELGIQUE; ESPAGNE; RESULTATS
- NOMS DIVERS : COUPE HOPMAN
NODOC: SR_20021231_475

Figure 3.2. Les informations en gris ont été supprimées manuellement avant de soumettre notre corpus à toute forme de traitement.

[3] Compte tenu de la nature de ces informations, nous sommes d'avis que le fait de les conserver aurait d'ailleurs augmenté la qualité des résultats obtenus.

3.2. Expérimentation 1 : application d'une technique prédictive

3.2.1. Paramètres d'expérimentation

Lors de notre première expérimentation, nous avons tout d'abord procédé à la segmentation de notre corpus initial. Pour cela, nous avons opté pour une segmentation par paragraphes, à raison d'un paragraphe par segment. Ce choix a été motivé par l'intuition selon laquelle la structure des paragraphes (que ce soit d'articles de journaux ou de tout autre genre) devrait normalement refléter la structure thématique des documents. En effet, le découpage d'un texte en paragraphes n'est pas le fruit du hasard; nous croyons qu'il dépend de l'organisation des idées ou des thèmes abordés dans un document. En segmentant notre corpus initial composé de 250 articles, nous avons obtenus un total de 2 181 paragraphes.

Par la suite, nous avons extrait le lexique de notre corpus et nous l'avons soumis à plusieurs opérations de filtrage visant à en diminuer substantiellement la taille. Ainsi, nous avons d'abord supprimé les mots fonctionnels du lexique initial. Pour ce faire, nous avons employé la liste de mots fonctionnels de la langue française disponible sur le site Internet du Professeur Jean Véronis[4]. Nous avons aussi appliqué un processus de lemmatisation. L'opération de lemmatisation a été réalisée en utilisant le logiciel de forage de textes WORDSTAT.

Au niveau du filtrage basé sur la fréquence d'apparition des mots dans notre corpus, nous avons supprimé, dans un premier temps, les mots dont la fréquence dans chacun des paragraphes était inférieure à 4. Par la suite, nous avons supprimé les mots dont la fréquence dans l'ensemble du corpus était inférieure à 15. Finalement, nous avons supprimé les mots figurant dans plus de 15% des paragraphes. Ces trois seuils ont été déterminés empiriquement, c'est-à-dire que nous avons effectué un survol du lexique. Suite à cette observation de surface, nous avons retenu les seuils jugés subjectivement optimaux nous permettant de filtrer le lexique. Suite à ces différentes opérations, le lexique épuré était constitué de 1 021 mots.

Afin de réduire encore davantage la taille du lexique et d'accroître les résultats obtenus lors des opérations subséquentes de catégorisation et de classification, nous avons procédé à

---

[4] www.up.univ-mrs.fr/~veronis/

un filtrage manuel des mots restant. Cette dernière opération de filtrage a été effectuée en calculant la valeur TF · IDF de chacun des mots. En nous basant sur cette mesure, nous avons décidé de ne retenir que les 88 mots dont la valeur TF · IDF était la plus élevée. Le lexique filtré retenu pour la constitution de la matrice était composé des mots (*features*) suivants (tableau 3.1) :

Arabe	Formation	Médical	Réseau
Attentat	Fromager	Mosquée	Santé
Cinéma	Gare	Musulman	Scientifique
Clijsters	Gouvernement	Musulmane	Set
Collège	Guerre	Nucléaire	SNCB
Commission	Henin	Numérique	Sport
Conseil	Informatique	Ordinateur	Sportif
Consommateur	Internet	Peine	Système
Création	Irak	Plainte	Technologie
Cuisine	Islamique	Police	Tennis
Directeur	Islamiste	Politique	Terrain
Docteur	Juge	Pomme	Tournoi
Droit	Justice	Porto	Train
Électrabel	Laboratoire	Prison	Tribunal
Électrique	Ligue	Procureur	Universitaire
Élève	Logiciel	Producteur	Université
Étude	Loi	Professeur	Vainqueur
Étudier	Malade	Professionnel	Vélo
Euros	Masters	Réalisateur	Victime
Fer	Match	Recette	Victoire
Film	Médecin	Religieux	Vin
Foi	Médecine	Religion	Violence

Tableau 3.1. Liste des mots retenus suite au filtrage du lexique[5].

Il est important de noter que les différentes opérations de filtrage, en plus de diminuer la taille du lexique initial (accélérant ainsi le temps de traitement des données et l'interprétation

[5] Comme nous l'avons mentionné, l'opération de lemmatisation n'a pas été précédée d'un marquage morphosyntaxique des données. Par conséquent, les résultats de la lemmatisation, bien que satisfaisants, sont imparfaits et approximatifs. Par exemple, on note dans la liste des mots retenus, la présence des mots « Musulman » et « Musulmane », lesquels, en raison de leur ambigüité syntaxique, n'ont pu être réduits à un même lemme.

ultérieure des résultats de la catégorisation et de la classification), ont aussi permis d'effectuer un filtrage des segments de notre corpus. En effet, plusieurs segments de notre corpus initial se sont avérés non-significatifs d'un point de vue thématique (segments très courts, segments obtenus en raison d'une erreur d'édition de la part des éditeurs de la version numérique du corpus, etc.). Ces segments n'étaient caractérisés par aucun terme thématique significatif. Traduits en vecteurs, ces segments auraient été entièrement composés de valeurs nulles. En filtrant le lexique initial, nous avons donc aussi filtré les segments dans lesquels ne figurait aucun des 88 termes retenus. Le filtrage du lexique a donc aussi permis de supprimer 556 segments (25,5%) thématiquement non-significatifs. Les traitements ultérieurs ont donc été appliqués sur 1 625 segments (composés chacun d'un seul paragraphe) caractérisés par la présence d'au moins un des 88 mots (*features*) retenus.

Après avoir procédé à la segmentation du corpus, à l'extraction et au filtrage du lexique et au filtrage des segments, nous avons traduit les segments retenus en une matrice de vecteurs pondérés (en fonction de la fréquence relative de chacun des mots dans chacun des vecteurs) de taille 1625 x 88.

À partir de cette matrice pondérée, nous avons mené deux expériences de catégorisation. Dans la première, nous avons divisé aléatoirement la matrice afin de générer les ensembles d'apprentissage et de test nécessaires à l'opération de catégorisation (le contenu de chacun des ensembles à été déterminé aléatoirement, en ne tenant pas compte de la distribution des catégories thématiques attribuées manuellement par les éditeurs du corpus). L'ensemble d'apprentissage a été constitué de 2/3 du corpus (1 083 segments), alors que l'ensemble de test a été constitué du 1/3 du corpus (542 segments). Comme en témoigne la littérature sur le problème de la catégorisation des données à l'aide de réseaux neuronaux, il s'agit d'un ratio classique pour ce genre de tâche. Les annexes 1 et 2 présentent les statistiques générées par l'application NEFCLASS-J pour l'ensemble d'apprentissage et l'ensemble de test (obtenus aléatoirement).

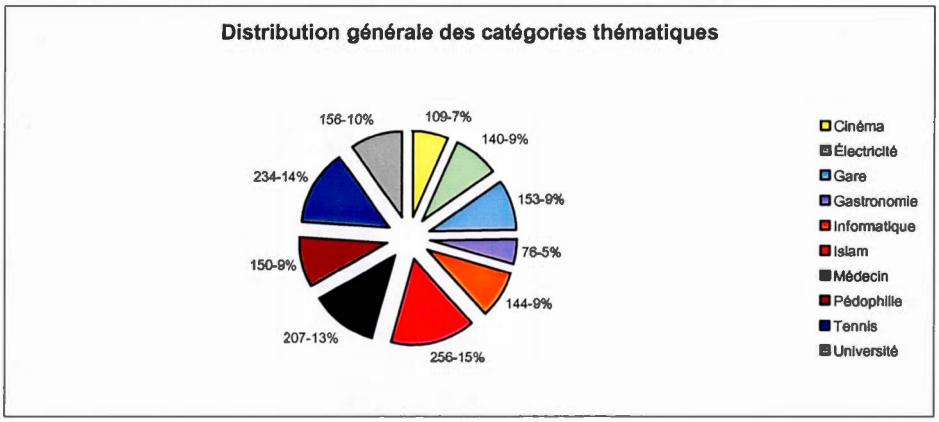


Figure 3.3. Distribution générale des catégories thématiques dans les 1 625 segments retenus.

Dans la seconde expérience de catégorisation, nous avons aussi découpé notre matrice selon le même ratio, mais nous nous sommes assuré, cette fois, de tenir compte de la distribution des catégories dans le corpus. Les annexes 3 et 4 présentent les statistiques générées par l'application NEFCLASS-J pour l'ensemble d'apprentissage et l'ensemble de test (obtenus en respectant la distribution des catégories).

3.2.2. Résultats obtenus lors de l'expérimentation 1

Les résultats obtenus grâce à l'opération de catégorisation automatique sont des regroupements de documents catégorisés. Nous avons aussi obtenus une liste de graphiques représentant la fuzzification de la fréquence de chacun des 88 mots retenus, ainsi que l'ensemble des règles (plus de 300 règles) automatiquement extraites ayant servi à la catégorisation des segments.

Les résultats de la catégorisation automatique ont été évalués en fonction de quatre mesures classiques dans le domaine du repérage de l'information et du forage de textes (Weiss *et al.*, 2005; Baeza-Yates et Ribeiro-Neto, 1999). Les indices utilisés sont les suivants : le taux de rappel, le taux de précision, le taux d'erreur et la mesure F (*F-Measure*) (Van Rijsbergen, 1979). Voici les formules décrivant chacune de ces mesures (formules 2, 3, 4 et 5) :

$$\text{Rappel} = \frac{a}{a + c} \quad (2)$$

$$\text{Précision} = \frac{a}{a + b} \quad (3)$$

$$\text{Taux d'erreur} = \frac{b + c}{a + b + c + d} \quad (4)$$

$$\textit{Mesure } F_\beta = \frac{(\beta^2 + 1)a}{(\beta^2 + 1)a + b + \beta^2 c} \quad (5)$$

Dans ces formules, *a* correspond au nombre de catégories attribuées tant par l'expert que par le système; *b* correspond au nombre de catégories attribuées par le système, mais non par l'expert; *c* correspond au nombre de catégories attribuées par l'expert, mais non par le système et *d* correspond au nombre de catégories non attribuées tant par l'expert que par le système.

Plusieurs études dans le domaine du repérage de l'information (Baeza-Yates et Ribeiro-Neto, 1999; Van Rijsbergen, 1979) ont démontré que les mesures de rappel et de précision, en plus de se heurter à d'importantes limites, peuvent varier de manière inversement proportionnelle. Van Rijsbergen (1979) a donc proposé une solution permettant de corriger cette situation en pondérant et en jumelant (en une seule mesure) les mesures de rappel et de précision. La mesure proposée par Van Rijsbergen porte le nom de Mesure-F (*F-Measure*). Dans cet mesure, F_0 nous permet d'obtenir le même résultat que le taux de précision, F_∞ nous permet d'obtenir le même résultat que le taux de rappel, F_1 nous permet d'accorder des poids équivalents au rappel et à la précision et $F_{0,5}$ nous permet d'accorder 2 fois plus d'importance au taux de précision qu'au taux de rappel. À des fins d'évaluation, nous avons employé dans la mesure F une valeur de 0,5.

Les tableaux suivants (3.2 et 3.3) présentent de manière synthétique les résultats obtenus par l'opération de catégorisation.

CATÉGORIE	MESURE	SCORE (1)	SCORE (2)
CINÉMA	Rappel	21,62	23,53
	Précision	50,00	50,00
	Mesure F (0,5)	39,60	40,82
	Taux d'erreur	6,83	6,27
ÉLECTRICITÉ	Rappel	39,53	44,74
	Précision	73,91	73,91
	Mesure F (0,5)	62,96	65,38
	Taux d'erreur	5,90	4,98
GARE	Rappel	60,00	69,23
	Précision	62,79	62,79
	Mesure F (0,5)	62,21	63,98
	Taux d'erreur	6,27	5,17
GASTRONOMIE	Rappel	31,03	33,33
	Précision	81,82	81,82
	Mesure F (0,5)	61,64	63,38
	Taux d'erreur	4,06	3,69
INFORMATIQUE	Rappel	14,81	16,67
	Précision	53,33	53,33
	Mesure F (0,5)	35,09	37,04
	Taux d'erreur	9,78	8,67
ISLAM	Rappel	46,15	48,84
	Précision	68,85	68,85
	Mesure F (0,5)	62,69	63,64
	Taux d'erreur	12,55	11,62
MÉDECIN	Rappel	64,79	70,77
	Précision	21,80	21,80
	Mesure F (0,5)	25,14	25,30
	Taux d'erreur	35,06	33,95
PÉDOPHILIE	Rappel	38,64	39,53
	Précision	54,84	54,84
	Mesure F (0,5)	50,60	50,90
	Taux d'erreur	7,56	7,38
TENNIS	Rappel	59,74	63,89
	Précision	93,88	93,88
	Mesure F (0,5)	84,25	85,82
	Taux d'erreur	6,27	5,35
UNIVERSITÉ	Rappel	49,02	51,02
	Précision	60,98	60,98
	Mesure F (0,5)	58,14	58,69
	Taux d'erreur	7,75	7,38
	Rappel moyen :	**42,53**	**46,15**
	Précision moyenne :	**62,22**	**62,22**
	Mesure F (0,5) moyenne :	**54,23**	**55,49**
	Taux d'erreur moyen :	**10,20**	**9,45**

Tableau 3.2. Évaluation des résultats de la catégorisation sur un corpus de test obtenu aléatoirement.

CATÉGORIE	MESURE	SCORE (1)	SCORE (2)
CINÉMA	Rappel	22,22	50,00
	Précision	50,00	50,00
	Mesure F (0,5)	40,00	50,00
	Taux d'erreur	6,65	2,96
ÉLECTRICITÉ	Rappel	36,17	58,62
	Précision	77,27	77,27
	Mesure F (0,5)	62,96	72,65
	Taux d'erreur	6,47	3,14
GARE	Rappel	66,67	82,93
	Précision	64,15	64,15
	Mesure F (0,5)	64,64	67,19
	Taux d'erreur	6,65	4,81
GASTRONOMIE	Rappel	40,00	55,56
	Précision	76,92	76,92
	Mesure F (0,5)	64,94	71,43
	Taux d'erreur	3,33	2,03
INFORMATIQUE	Rappel	12,50	24,00
	Précision	40,00	40,00
	Mesure F (0,5)	27,78	35,29
	Taux d'erreur	9,43	5,18
ISLAM	Rappel	56,47	87,27
	Précision	67,61	67,61
	Mesure F (0,5)	65,04	70,80
	Taux d'erreur	11,09	5,55
MÉDECIN	Rappel	28,99	48,78
	Précision	48,78	48,78
	Mesure F (0,5)	42,92	48,78
	Taux d'erreur	12,94	7,76
PÉDOPHILIE	Rappel	38,00	51,35
	Précision	55,88	55,88
	Mesure F (0,5)	51,08	54,91
	Taux d'erreur	8,50	6,10
TENNIS	Rappel	58,97	75,41
	Précision	93,88	93,88
	Mesure F (0,5)	83,94	89,49
	Taux d'erreur	6,47	3,33
UNIVERSITÉ	Rappel	53,85	70,00
	Précision	57,14	57,14
	Mesure F (0,5)	56,45	59,32
	Taux d'erreur	8,32	6,10
	Rappel moyen :	**41,38**	**60,39**
	Précision moyenne :	**63,16**	**63,16**
	Mesure F (0,5) moyenne :	**55,97**	**61,99**
	Taux d'erreur moyen :	**7,99**	**4,70**

Tableau 3.3. Évaluation des résultats de la catégorisation sur un corpus de test respectant la distribution initiale des catégories.

3.2.3. Discussion des résultats de l'expérimentation 1

Avant de commenter ces résultats, il importe de mentionner que dans certaines situations l'application NEFCLASS-J n'est pas en mesure de générer un ensemble de règles qui permettrait de catégoriser l'ensemble des données disponibles. Ceci est attribuable aux processus de généralisation et d'élagage des règles impliqués dans la démarche. L'application génère donc un ensemble de règles généralisées et élaguées permettant de couvrir le maximum de cas possibles. L'application trouve donc un équilibre entre, d'une part, le nombre de règles extraites et, d'autre part, l'efficacité effective des règles. Par ailleurs, il existe aussi des cas pour lesquels il est impossible de trouver une règle de catégorisation qui ne serait pas en contradiction avec une autre règle (dont l'efficacité est déjà attestée). Par conséquent, lors de certaines expérimentations, l'application se trouve dans l'impossibilité de catégoriser certains cas. Dans le cadre de nos travaux, nous avons été confronté à cette limite des systèmes fondés sur l'extraction de règles. Nous avons donc évalué les résultats de la catégorisation de deux manière, l'une en tenant compte des cas non-catégorisés (Score (1)), l'autre en ignorant ces cas problématiques (Score (2)).

En consultant les résultats figurant dans les tableaux 3.2 et 3.3, on constate que le processus de catégorisation automatique employant une technique hybride neuro-floue a donné des résultats plutôt décevants. Pour le corpus de test ne tenant pas compte de la distribution des catégories, lorsque l'on tient compte des segments que le système n'a pas catégorisés (que l'on a alors considérés comme étant mal catégorisés), les performances moyennes du système sont les suivantes : le taux de rappel est de 42,53%, le taux de précision est de 62,22%, la mesure F (0,5) est de 54,23% et le taux d'erreur est de 10,20%. Si l'on ne tient pas compte des cas ignorés, les résultats sont très légèrement supérieurs. Ainsi, dans ce contexte, le taux de rappel est de 46,15%, le taux de précision est de 62,22%, la mesure F (0,5) est de 55,49% et le taux d'erreur est de 9,45%.

Pour le corpus de test tenant compte de la distribution des catégories, lorsque l'on tient compte des segments que le système n'a pas catégorisés (que l'on a alors considérés comme étant mal catégorisés), les performances moyennes du système sont les suivantes : le taux de rappel est de 41,38%, le taux de précision est de 63,16%, la mesure F (0,5) est de 55,97% et le taux d'erreur est de 7,99%. Si l'on ne tient pas compte des cas ignorés, les résultats sont

très légèrement supérieurs. Ainsi, dans ce contexte, le taux de rappel est de 60,39%, le taux de précision est de 63,16%, la mesure F (0,5) est de 61,99% et le taux d'erreur est de 4,70%.

La première observation à mentionner concernant ces résultats est la suivante : le fait de tenir ou non compte des cas (segments) ignorés par le système n'a un impact significatif que sur le taux de rappel. On constate à cet égard une amélioration d'environ 20%. Cependant, que l'on tienne compte ou non de ces documents le taux de précision reste identique. Il est par contre intéressant de noter que le fait de constituer les corpus d'apprentissage et de test en tenant compte ou non de la distribution des catégories n'influence pas significativement la qualité des résultats obtenus.

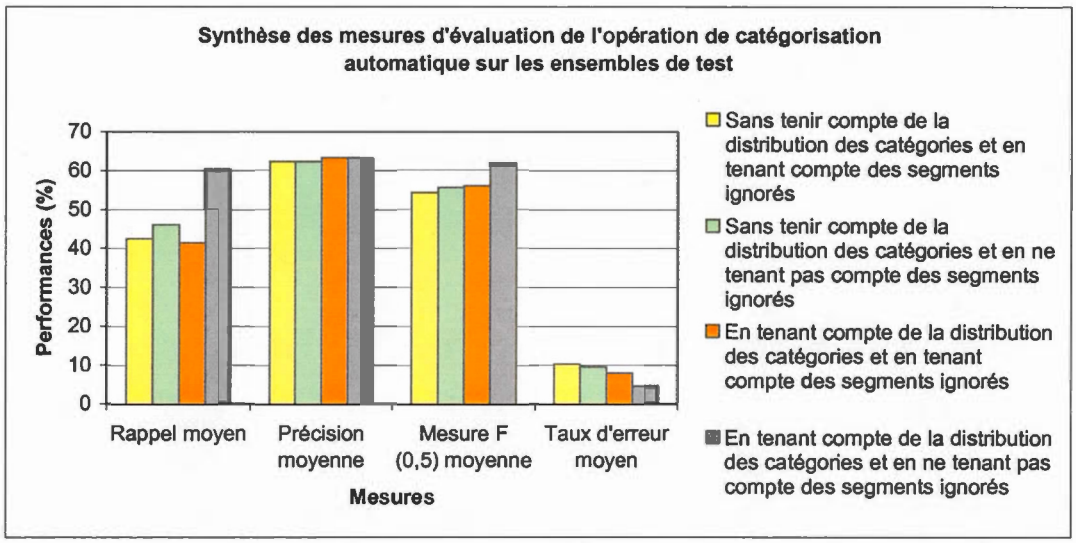


Figure 3.4. Synthèse des mesures d'évaluation de l'opération de catégorisation automatique sur les ensembles de test.

Lors d'expérimentations préliminaires (sur un corpus composé de 100 articles de journaux provenant de deux catégories thématiques) sur la base desquels nous avons choisi le modèle de catégorisation hybride neuro-flou (Forest, 2005), nous avions obtenus des résultats parfaits avec l'application NEFCLASS-J. Il importe donc d'identifier les raisons pour lesquelles les performances du système se sont détériorées lorsque nous l'avons appliqué sur un corpus de même genre, mais de taille largement supérieure.

Une première observation permet d'expliquer les faibles performances du système.

Comme nous l'avons mentionné précédemment, notre démarche méthodologique est fondée en partie sur la segmentation des documents. Nous croyons qu'une telle opération permet d'attribuer facilement plusieurs étiquettes ou catégories thématiques à un même document (ce qui est manifestement plus difficile à réaliser en traitant chaque document en entier). Cependant, afin d'évaluer les résultats de la catégorisation, nous nous sommes basé sur le travail de catégorisation manuelle effectué par les éditeurs du journal LE SOIR. Or, cette catégorisation manuelle a été effectuée en attribuant une ou plusieurs catégories thématiques à chaque document entier. Afin de ne pas biaiser notre évaluation, nous avons respecté le travail effectué manuellement et nous avons donc attribué la catégorie des documents à l'ensemble des paragraphes constitutifs de ces mêmes documents. Sans remettre totalement en question la qualité du travail effectué manuellement, nous sommes cependant en droit de constater que la (les) catégorie(s) attribuée(s) manuellement à chaque document ne s'applique(nt) peut-être pas à l'ensemble des paragraphes constituant ces documents. Nous aurions pu catégoriser manuellement chacun des paragraphes, mais cette opération préliminaire aurait fait intervenir une composante subjective qui aurait remis en doute la rigueur de la démarche employée. Par ailleurs, notre démarche présuppose que le travail effectué manuellement a été rigoureusement effectué. Une observation minutieuse du corpus analysé nous indique que le travail manuel n'a pas toujours été effectué avec la rigueur souhaitée. Ces premières observations peuvent expliquer, du moins en partie, les faibles performances du système.

Nous notons aussi qu'un autre facteur peut expliquer les performances du système. Celui-ci relève du critère de segmentation du corpus. Ainsi, plusieurs études ont tenté d'identifier la taille optimale que doivent avoir les différents segments d'un corpus (Callan, 1994; Moffat *et al.*, 1994; Hearst et Plaunt, 1993). Plusieurs de ces études indiquent que la segmentation par paragraphes, en plus d'être intuitivement attrayante, possède de nombreux avantages. On constate, par contre, que les articles journalistiques sont composés de paragraphes dont la taille est très petite (quelques dizaines de mots, dans bien des cas). Ceci n'est pas sans engendrer de nombreux problèmes lorsqu'il s'agit de traiter informatiquement des données textuelles. En effet, plus les segments sont petits, plus il devient difficile d'y retrouver des mots thématiquement significatifs qu'il est aussi possible de retrouver dans d'autres segments traitant d'un même thème. En d'autres termes, on constate que plus la taille des segments est

petite, plus la variation lexicale nécessaire afin de différentier deux segments sera petite. Bref, cela revient à dire que, lorsque nous comparons des segments de petite taille, une faible variation lexicale d'un ou deux mots s'avère suffisante pour catégoriser des segments thématiquement reliés dans des catégories thématiques différentes. Par conséquent, il est possible que, dans le cas d'une tâche de catégorisation automatique d'articles de journaux, la segmentation par paragraphes (à raison d'un paragraphe par segment) ne soit pas optimale.

En sommes, selon les modalités d'expérimentation spécifiées et les résultats obtenus dans notre projet, l'hypothèse selon laquelle une méthodologie fondée sur la catégorisation automatique des segments de documents peut permettre d'assister l'identification du contenu thématique et l'analyse thématique est donc infirmée.

Doit-on pour autant rejeter le modèle de catégorisation hybride neuro-flou pour le traitement des documents textuels? Bien que les résultats obtenus lors de nos expérimentations se soient avérés inférieurs à nos attentes, nous sommes d'avis que le modèle de catégorisation que nous avons employé peut néanmoins s'avérer utile à des fins d'analyse des documents textuels non-structurés.

Afin d'optimiser les performances de ce système, il nous sommes d'avis qu'importe d'explorer davantage les paramètres sur lesquels reposent en grande partie les performances du système. À cet égard, nous croyons qu'il est possible d'accroître considérablement les performances du système en faisant varier plusieurs paramètres, tels que la taille des segments et le nombre de variables (*features*) constituant les vecteurs soumis au système, ou en intégrant un thésaurus permettant d'enrichir la liste de variables retenues.

Par ailleurs, plusieurs études ont récemment démontré la pertinence de différentes méthodes de catégorisation automatique à des fins d'analyse de données textuelles. Les performances obtenues lors de nos expérimentations ne peuvent, à elles seules, remettre en question les conclusions démontrées par d'autres études dans le même domaine. Au niveau théorique, l'utilisation de processus de catégorisation automatique n'est plus à remettre en question dans le domaine de l'analyse et de la gestion des documents textuels. Plutôt que de remettre en question cette utilisation, nous croyons qu'il importe davantage de s'interroger sur les modalités d'application qui pourraient permettre d'accroître les performances de ces techniques.

3.3. Expérimentation 2 : application d'une technique exploratoire

3.3.1. Paramètres d'expérimentation

Lors de notre seconde expérimentation, nous avons appliqué les mêmes opérations de prétraitement et de filtrage que celles qui furent appliquées lors de la première expérimentation. Ainsi, notre corpus à été segmenté en paragraphes (à raison d'un paragraphe par segment). Les mots fonctionnels ont été supprimés du lexique initial et un processus de lemmatisation à été appliqué. Nous avons supprimé les mots dont la fréquence dans chacun des paragraphes était inférieure à 4, ceux dont la fréquence dans l'ensemble du corpus était inférieure à 15 et ceux figurant dans plus de 15% des paragraphes. Finalement, nous avons retiré manuellement les termes dont la valeur TF · IDF a été empiriquement jugée trop faible, pour ne retenir, en dernière instance, que les mêmes 88 mots retenus lors de l'expérimentation précédente. Suite aux opérations de segmentation et de filtrage, nous avons vectorisé les segments et généré une matrice de taille 1625 x 88. Pour notre seconde expérimentation, nous avons cependant généré une matrice binaire ne tenant pas compte de la fréquence d'apparition de chacun des mots dans chacun des segments. Ce choix nous a été dicté par le classifieur auquel nous avons soumis la matrice. En effet, nous avons choisi d'utiliser le classifieur neuronal ART1, lequel n'accepte en intrant que des vecteurs binaires.

Par la suite, l'ensemble de nos vecteurs ont été classifiés. L'opération de classification a été effectuée selon les paramètres suivants : le paramètre de vigilance de l'algorithme a été fixé à 0,005 et le nombre d'itérations de l'algorithme de classification a été fixé à 2 500.

En tenant compte de ces paramètres, l'opération de classification a regroupé les 1 625 vecteurs en 118 classes. Nous avons donc obtenu une moyenne de 13,77 segments par classes. Nous avons privilégié une classification très fine (rappelons que dans ART1, plus la valeur du paramètre de vigilance est faible, plus la classification effectuée est fine et détaillée), car nous avons voulu identifier le plus fidèlement possible les distinctions et les variations thématiques observables dans le corpus. Nous avons aussi voulu explorer si une telle approche nous permettrait d'identifier de fines variations thématiques dont les éditeurs du corpus n'ont pu (ou n'ont voulu) tenir compte lors de leur catégorisation manuelle. La figure 3.5 présente la distribution des catégories thématiques dans chacune des classes (les

représentations de la distribution des catégories thématiques dans chaque classe individuelle peuvent être consultées dans l'annexe 5).

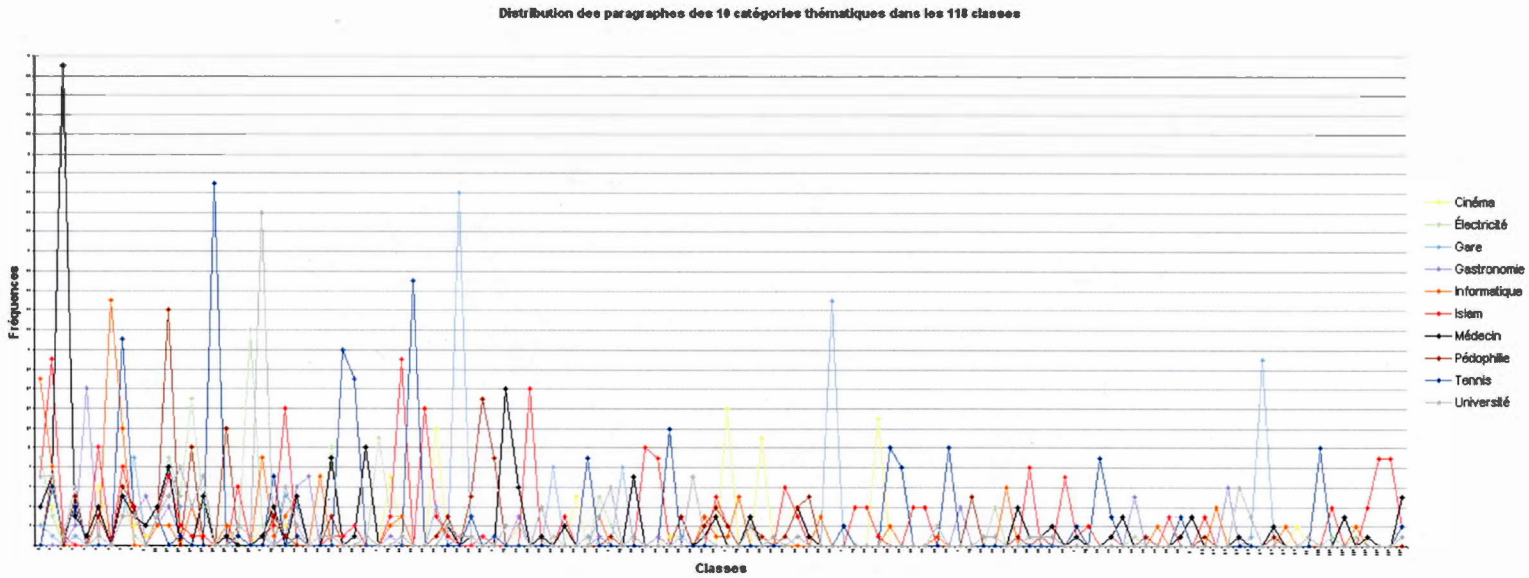


Figure 3.5. Distribution des paragraphes des 10 catégories thématiques dans les 118 classes.

3.3.2. Résultats obtenus lors de l'expérimentation 2

Comme nous l'avons mentionné précédemment, la classification est une opération non-supervisée de nature exploratoire. Elle nous permet d'identifier des regroupements – de nature thématique, comme nous le verrons dans les pages suivantes – sans faire intervenir des métadonnées ou des connaissances (externes) de l'utilisateur. Elle permet donc, indépendamment de ce dernier, d'extraire de précieuses informations sur l'organisation et la structure des données. Dans une telle perspective exploratoire, l'évaluation des résultats de la classification selon les méthodes classiques dans le domaine du repérage de l'information est pratiquement impossible. Les résultats de l'opération de classification ne peuvent pas être comparés à des résultats de référence. Ils ne font que témoigner d'une structure manifeste qui peut être observée dans les données. Par conséquent, notre objectif n'est pas d'évaluer les résultats de la classification[6], mais plutôt ceux de l'opération qui l'accompagne dans le cadre de ce projet, à savoir l'identification automatique des termes thématiques candidats.

Afin d'identifier le contenu thématique des classes, nous avons choisi d'appliquer deux mesures spécifiques permettant d'extraire automatiquement à partir du lexique de chaque classes les termes candidats pouvant servir d'étiquette thématique. Les deux mesures utilisées sont la fréquence relative et la valeur TF · IDF. Ainsi, nous avons donc retenu comme termes candidats de chaque classe les 3 termes possédant la fréquence et la valeur TF · IDF la plus élevée. Avant de présenter la liste des termes thématiques candidats retenus pour chaque classe, il importe de mentionner que les deux mesures employées (la fréquence et la fréquence inverse (TF · IDF)) ont généré des listes de termes candidats presque identiques. Pour cette tâche, il est donc possible d'employer l'une ou l'autre mesure sans affecter significativement les résultats. Le tableau suivant (tableau 3.4) présente la liste des termes retenus pour chaque classe.

---

[6] L'adéquation et la pertinence de ces résultats sont jugées ultimement par l'utilisateur, en regard d'un usage possible.

CLASSE	MOTS THÉMATIQUES CANDIDATS
001	Système, Microsoft, Ordinateur
002	Avocat, Ministre, Politique
003	Médecin, Généraliste, Patient
004	Professeur, Festival, Classe
005	Produit, Table, Goût
006	Violence, Communauté, Film
007	Logiciel, Microsoft, Sécurité
008	Foi, Femme, Énergie
009	Chemin [de fer], Police, Gare
010	Recette, Cuisine, Table
011	Porto, Avocat, Cuisine
012	Dutroux, Dossier, Nihoul
013	Création, Société, Réseau,
014	Réseau, Électricité, Électrabel,
015	Étude, Énergie, Secondaire,
016	Tournoi, Clijsters, Henin
017	Plainte, Dossier, Procureur
018	Attentat, Musulman, États-Unis
019	Électrique, Vélo, Transport
020	Étudier, Enseignement, Université
021	Victoire, Clijsters, Débat
022	Politique, Pouvoir, Communauté
023	Coût, Projet, Prix
024	Maison, Chef, Produit
025	Informatique, Société, Projet
026	Commission, Loi, Enseignement
027	Henin, Clijsters, Williams
028	Victoire, Williams, Tennis
029	Médical, UCL, Santé
030	Électrabel, Électricité, Producteur
031	Producteur, Film, Jeu
032	Gouvernement, Ministre, Islamiste
033	Clijsters, Balle, Jeu
034	Musulman, Communauté, Ministre
035	Cinéma, Industrie, Studio
036	Directeur, Projet, Enfant
037	Gare, Quartier, Voie
038	Professionnel, Avocat, Dossier
039	Procureur, Avocat, Bourlet
040	Tribunal, Avocat, Juge
041	Médecin, Médical, Patient
042	Santé, Soin, Spécialiste
043	Musulman, Mosquée, Islamiste
044	Enseignement, Université, Recherche
045	Train, Vie, Projet
046	Arabe, Monde, Valeur
047	Acteur, Monde, Confiance

CLASSE	MOTS THÉMATIQUES CANDIDATS
048	Sport, Tennis, Stade
049	Consommateur, Compte, Monde
050	Laboratoire, ULB, Recherche
051	SNCB, Passage, Collège
052	Santé, Politique, Soin
053	Musulman, Politique, Monde
054	Religieux, AKP, Média
055	Set, Match, Jeu
056	Collège, Procès, Enfant
057	Université, Débat, Formation
058	Internet, Informatique, Page
059	Justice, International, Équipe
060	Film, Scène, Image
061	Logiciel, Réseau, Virus
062	Professeur, Recherche, Islamiste
063	Film, Scène, Numérique
064	Vélo, Sécurité, Qualité,
065	Police, Quartier, Fédéral
066	Loi, Enseignement, Enfant
067	Victime, Autorité, Dossier
068	Technologie, Entreprise, Secteur
069	Gare, Train, SNCB
070	Masters, Gagner, Mondial
071	Mosquée, Quartier, Turc
072	Musulman, Islam, Communauté
073	Film, Cinéma, Scène
074	Match, Jeu, Joueur
075	Henin, Match, Coup
076	Islamiste, Ville, Coup
077	Islamique, Art, Projet
078	Formation, Organisation, Université
079	Victoire, Tournoi, Coup
080	Cuisine, Restaurer, Chef
081	Juge, Avocat, Client
082	Université, Étude, Association
083	Nucléaire, Réaction, Produit
084	Informatique, Réseau, Technologie
085	Prison, Vie, Malade
086	Gouvernement, Politique, Chef
087	Scientifique, Machine, Université
088	Malade, Vie, Patient
089	Arabe, Politique, Monde
090	Vainqueur, Femme, Vie
091	Religion, Monde, Islam
092	Tennis, Saison, Coupe
093	Terrain, Terre, Territoire
094	Docteur, Cas, Bourgmestre

CLASSE	MOTS THÉMATIQUES CANDIDATS
095	Pomme, Fruit, Produit
096	Dénoncer, Chef, Judiciaire
097	Travailler, Ordinateur, Texte
098	Politique, Islamique, Religieux
099	Sport, Idée, Résultat
100	Médecin, Étude, Santé
101	Droit, Gouvernement, Ministre
102	Ordinateur, Informatique, Sécurité
103	Menu, Carte, Table
104	Étudier, Professeur, Programme
105	Foi, Dossier, Recherche
106	Gare, SNCB, Parking
107	Victime, Plainte, Médecin
108	Numérique
109	Scène, Guerre, Passage
110	Magistrat, Police, Justice
111	Clijsters, Masters, Henin
112	Conseil, Gouvernement, Ministre
113	Scientifique, Médical, Famille
114	Système, Informatique, Client
115	Droit, Foi, Pouvoir
116	Musulman, Religion, Islam
117	Arabe, Ligue, Islamiste
118	Médecine, Recherche, Médical

Tableau 3.4. Liste des termes thématiques candidats de chaque classe (3 termes par classe).

L'évaluation des résultats obtenus lors de la classification automatique et de l'extraction automatique des termes thématiques candidats ne saurait être réalisée à l'aide des mesures employées pour évaluer les résultats obtenus lors de l'expérimentation de catégorisation. En effet, l'utilisation de mesures telles que le taux de rappel et le taux de précision présuppose que nous disposions de résultats de référence (*benchmark*) à partir desquels il est possible de comparer les résultats obtenus. Compte tenu de la variété de termes présents dans chacune des classes thématiques, il est très difficile d'évaluer avec exactitude la pertinence des termes extraits pouvant être considérés à juste titre comme étiquette thématique valide. Par ailleurs, en l'absence de résultats gabarits, la subjectivité inhérence au processus de thématisation implique que certains termes retenus sont pertinents ou non en fonction des attentes et des objectifs de l'utilisateur.

> Mais comment la pertinence ou la non-pertinence thématique d'une unité est-elle déterminée? En tant que thématiseur, je choisis ou construit un cadre thématique (comportant un nombre indéfini d'autres cadres) en fonction duquel une grande quantité d'unités peuvent former un tout. Ce cadre est (fondé sur) un modèle qui dérive d'une réalité intra- ou extratextuelle (et la sélection comme constitution en sont conditionnées par mon savoir, mes intérêts et mes buts). (Prince, 1985, p. 430)

Malgré la subjectivité du processus de thématisation, ce n'est manifestement pas n'importe quel terme qui peut être considéré comme étiquette thématique représentant le contenu de la classe à laquelle il appartient. En effet, comme nous l'avons mentionné, tout texte possède une contrainte objective qui entre aussi en compte dans le cadre du processus de thématisation. Cette contrainte repose sur le texte à analyser. Cette composante intra-textuelle limite nécessairement la liberté de l'interprète. C'est elle qui guide – du moins partiellement – l'ensemble des analyses. Ainsi, malgré les intérêts, les connaissances et les objectifs du chercheur, ce dernier ne crée pas les thèmes dans le corpus qu'il analyse. C'est le texte qui expose, à l'aide des différents porteurs sémiotiques qu'il comporte, les thèmes sur lesquels le chercheur posera éventuellement son analyse.

Il importe, dès lors, de prendre en considération ces deux perspectives dans l'évaluation des résultats du processus d'extraction automatique des termes thématiques. Le défi de cette tâche d'évaluation réside donc dans le compromis entre d'une part, la composante subjective du processus interprétatif et, d'autre part, les marqueurs linguistiques composant les documents soumis à l'analyse. L'évaluation nous apparaît donc des plus complexes.

Afin d'évaluer les résultats de la classification et de l'identification automatique des termes thématiques candidats, nous avons utilisé les réseaux lexicaux provenant de la base de données informatisée WORDNET (Fellbaum, 1998; Miller *et al.*, 1993). Cette base de données est inspirée de recherches récentes en psycholinguistique sur la mémoire lexicale. Elle consiste en plusieurs réseaux de lexèmes où chaque nœud correspond à un concept et est représenté par un ensemble de mots (constituant ainsi un *synset*). On y retrouve aussi une brève définition (*gloss*) de chacun des termes présents dans la base de données. Les *synsets* sont divisés en quatre principales catégories syntaxiques (nom, verbe, adjectif et adverbe). À partir d'une requête simple composée d'un seul terme, la principale fonctionnalité de cette application est de relier le terme de la requête initiale (figurant dans un *synset*) aux différents termes (figurant dans d'autres *synsets*) qui lui sont sémantiquement associés selon certaines

relations spécifiques. Ainsi, il est possible d'obtenir les différents termes partageant l'une ou l'autre relation sémantique avec le terme initial (figure 3.6). Plusieurs relations générant les réseaux sémantiques peuvent être explorées. Parmi celles-ci, on retrouve, entre autres, les relations synonymique, holonymique, hyperonymique, hyponymique, méronymique, etc.

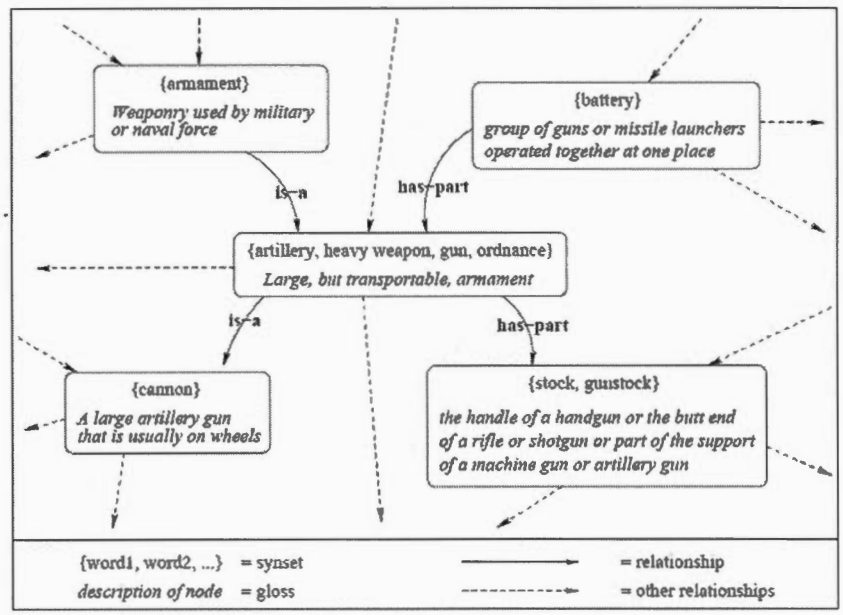


Figure 3.6. Représentation des *synsets* et des relations dans WORDNET (tiré de Patwardhan, 2003, p.9).

Pour effectuer une telle évaluation, nous devions connaître *a priori* la véritable catégorie thématique de chacune des classes. Pour ce faire, nous avons décidé d'attribuer comme véritable étiquette thématique de chaque classe le nom de la catégorie thématique des segments constituant le plus grand pourcentage de la classe. À titre d'exemple, nous avons attribué à la classe 1 la catégorie « INFORMATIQUE », car elle est constituée à 35% de segments provenant de documents catégorisés par cette étiquette (voir figure 3.7).

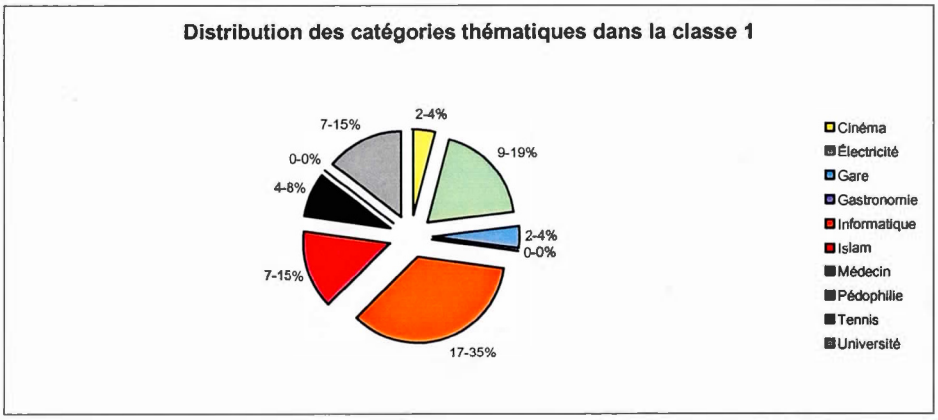


Figure 3.7. Distribution des catégories thématiques dans la classe 1.

Dans les quelques rares cas d'égalité, où plus d'une catégorie représentait, en termes de pourcentages, le contenu d'une même classe, nous avons choisi d'attribuer autant de catégories que nécessaire aux classes concernées. Par exemple, nous avons attribué à la classe 64 les étiquettes thématiques « ÉLECTRICITÉ » et « UNIVERSITÉ » (figure 3.8).

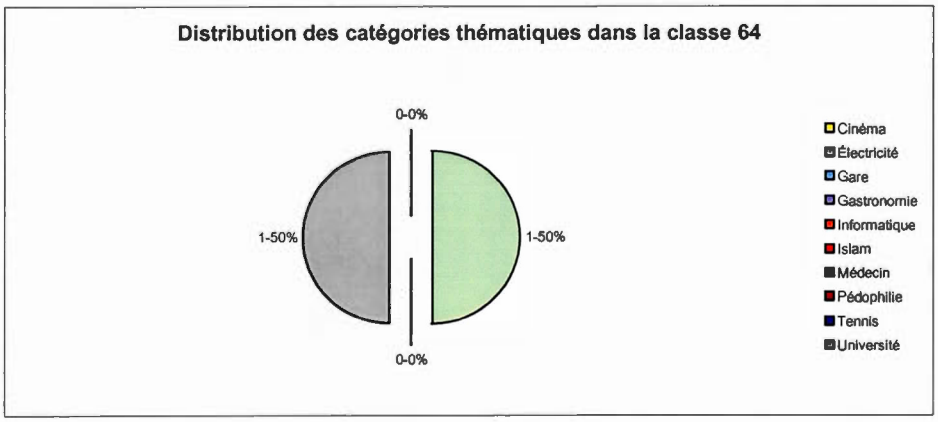


Figure 3.8. Distribution des catégories thématiques dans la classe 64.

Nous avons donc évalué la proximité sémantique entre la ou les catégories réelles d'une classe de segments et les termes candidats pouvant servir d'étiquette thématique. Les termes candidats ont été systématiquement comparés à ceux composant le réseau lexical proposé par WORDNET pour cette même catégorie. Si, selon les données de WORDNET, les termes

candidats sont reliés (en fonction de relations pertinentes, telles la synonymie ou l'hyponymie) aux catégories réelles de la classe à laquelle ils appartiennent, alors nous pourrons les considérer comme de bonnes étiquettes thématiques représentant le contenu de la classe à laquelle ils appartiennent.

La version française de la base de données lexicale WORDNET n'étant pas disponible gratuitement (elle est même assez dispendieuse), nous avons dû effectuer la traduction des dix catégories initiales, ainsi que de l'ensemble des termes thématiques candidats extraits avant d'effectuer l'évaluation des résultats. Les tableaux 3.5 et 3.6 présentent les termes français employés, ainsi que leur équivalent dans la langue anglaise.

CATÉGORIES (TERMES FRANÇAIS)	CATÉGORIES (TERMES ANGLAIS)
Cinéma	*Cinema*
Électricité	*Electricity*
Gare	*Train station*
Gastronomie	*Gastronomy*
Informatique	*Computer, computer science*
Islam	*Islam*
Médecin	*Doctor, physician*
Pédophilie	*Pedophilia*
Tennis	*Tennis*
Université	*University*

Tableau 3.5. Traduction des catégories initiales.

CLASSE	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	MOTS THÉMATIQUES CANDIDATS (ANGLAIS)
001	Système, Microsoft, Ordinateur	*System, Microsoft, Computer*
002	Avocat, Ministre, Politique	*Lawyer, Minister, Politics*
003	Médecin, Généraliste, Patient	*Doctor, Family doctor, Patient*
004	Professeur, Festival, Classe	*Professor, Festival, Class*
005	Produit, Table, Goût	*Product, Table, Taste*
006	Violence, Communauté, Film	*Violence, Community, Movie, film*
007	Logiciel, Microsoft, Sécurité	*Software, Microsoft, Security*
008	Foi, Femme, Énergie	*Faith, Woman, Energy*
009	Chemin [de fer], Police, Gare	*Railway, Police, Train station*
010	Recette, Cuisine, Table	*Recipe, Cuisine, Table*
011	Porto, Avocat, Cuisine	*Porto, Lawyer, Cuisine*
012	Dutroux, Dossier, Nihoul	*Dutroux, File, Nihoul*
013	Création, Société, Réseau	*Creation, Society, Network*
014	Réseau, Électricité, Électrabel	*Network, Electricity, Électrabel*
015	Étude, Énergie, Secondaire	*Study, Energy, High-school*
016	Tournoi, Clijsters, Henin	*Tournament, Clijsters, Henin*

CLASSE	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	MOTS THÉMATIQUES CANDIDATS (ANGLAIS)
017	Plainte, Dossier, Procureur	*Complaint, File, Prosecutor*
018	Attentat, Musulman, États-Unis	*Attack, Muslim, United States*
019	Électrique, Vélo, Transport	*Electric, Bicycle, Transportation*
020	Étudier, Enseignement, Université	*Study, Teaching, University*
021	Victoire, Clijsters, Débat	*Victory, Clijsters, Debate*
022	Politique, Pouvoir, Communauté	*Politics, Power, Community*
023	Coût, Projet, Prix	*Cost, Project, Price*
024	Maison, Chef, Produit	*Home made, Chef, Product*
025	Informatique, Société, Projet	*Computer, Society, Project*
026	Commission, Loi, Enseignement	*Commission, Law, Teaching*
027	Henin, Clijsters, Williams	*Henin, Clijsters, Williams*
028	Victoire, Wiiliams, Tennis	*Victory, Williams, Tennis*
029	Médical, UCL, Santé	*Medical, UCL, Health*
030	Électrabel, Électricité, Producteur	*Électrabel, Electricity, Producer*
031	Producteur, Film, Jeu	*Producer, Movie, film, Play, acting*
032	Gouvernement, Ministre, Islamiste	*Government, Minister, Islamic*
033	Clijsters, Balle, Jeu	*Clijsters, Ball, Game*
034	Musulman, Communauté, Ministre	*Muslim, Community, Minister*
035	Cinéma, Industrie, Studio	*Cinema, Industry, Studio*
036	Directeur, Projet, Enfant	*Director, Project, Child*
037	Gare, Quartier, Voie	*Train station, Neighbourhood, Lane*
038	Professionnel, Avocat, Dossier	*Professional, Lawyer, File*
039	Procureur, Avocat, Bourlet	*Prosecutor, Lawyer, Bourlet*
040	Tribunal, Avocat, Juge	*Court, Lawyer, Judge*
041	Médecin, Médical, Patient	*Doctor, Medical, Patient*
042	Santé, Soin, Spécialiste	*Health, Care, Specialist*
043	Musulman, Mosquée, Islamiste	*Muslim, Mosque, Islamist*
044	Enseignement, Université, Recherche	*Teaching, University, Research*
045	Train, Vie, Projet	*Train, Life, Project*
046	Arabe, Monde, Valeur	*Arabic, World, Value*
047	Acteur, Monde, Confiance	*Actor, World, Confidence*
048	Sport, Tennis, Stade	*Sport, Tennis, Stadium*
049	Consommateur, Compte, Monde	*Consumer, Account, World*
050	Laboratoire, ULB, Recherche	*Laboratory, ULB, Research*
051	SNCB, Passage, Collège	*SNCB, Way, College*
052	Santé, Politique, Soin	*Health, Politics, Care*
053	Musulman, Politique, Monde	*Muslim, Politics, World*
054	Religieux, AKP, Média	*Religious, AKP, Media*
055	Set, Match, Jeu	*Set, Match, Game*
056	Collège, Procès, Enfant	*College, Trial, Child*
057	Université, Débat, Formation	*University, Debate, Training*
058	Internet, Informatique, Page	*Internet, Computer, Page*
059	Justice, International, Équipe	*Justice, International, Team*
060	Film, Scène, Image	*Movie, film, Scene, Image*
061	Logiciel, Réseau, Virus	*Software, Network, Virus*
062	Professeur, Recherche, Islamiste	*Professor, Research, Islamist*

CLASSE	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	MOTS THÉMATIQUES CANDIDATS (ANGLAIS)
063	Film, Scène, Numérique	*Movie, film, Scene, Digital*
064	Vélo, Sécurité, Qualité,	*Bicycle, Security, Quality,*
065	Police, Quartier, Fédéral	*Police, Neighbourhood, Federal*
066	Loi, Enseignement, Enfant	*Law, Teaching, Child*
067	Victime, Autorité, Dossier	*Victim, Authority, File*
068	Technologie, Entreprise, Secteur	*Technology, Business, Sector*
069	Gare, Train, SNCB	*Train station, Train, SNCB*
070	Masters, Gagner, Mondial	*Masters, Win, World*
071	Mosquée, Quartier, Turc	*Mosque, Neighbourhood, Turk*
072	Musulman, Islam, Communauté	*Muslim, Islam, Community*
073	Film, Cinéma, Scène	*Movie, film, Cinema, Scene*
074	Match, Jeu, Joueur	*Match, Game, Player*
075	Henin, Match, Coup	*Henin, Match, Hit*
076	Islamiste, Ville, Coup	*Islamist, City, Hit*
077	Islamique, Art, Projet	*Islamic, Art, Project*
078	Formation, Organisation, Université	*Training, Organisation, University*
079	Victoire, Tournoi, Coup	*Victory, Championship, Hit*
080	Cuisine, Restaurer, Chef	*Cuisine, Eat, Chef*
081	Juge, Avocat, Client	*Judge, Lawyer, Client*
082	Université, Étude, Association	*University, Study, Association*
083	Nucléaire, Réaction, Produit	*Nuclear, Reaction, Product*
084	Informatique, Réseau, Technologie	*Computer, Network, Technology*
085	Prison, Vie, Malade	*Jail, Life, Patient*
086	Gouvernement, Politique, Chef	*Government, Politics, Leader*
087	Scientifique, Machine, Université	*Scientific, Machine, University*
088	Malade, Vie, Patient	*Patient, Life, Patient*
089	Arabe, Politique, Monde	*Arabic, Politics, World*
090	Vainqueur, Femme, Vie	*Winner, Woman, Life*
091	Religion, Monde, Islam	*Religion, World, Islam*
092	Tennis, Saison, Coupe	*Tennis, Season, Cup*
093	Terrain, Terre, Territoire	*Court, Clay, Territory*
094	Docteur, Cas, Bourgmestre	*Doctor, Case, Bourgmestre*
095	Pomme, Fruit, Produit	*Apple, Fruit, Product*
096	Dénoncer, Chef, Judiciaire	*Denounce, Leader, Judicial*
097	Travailler, Ordinateur, Texte	*Work, Computer, Text*
098	Politique, Islamique, Religieux	*Politics, Islamic, Religious*
099	Sport, Idée, Résultat	*Sport, Idea, Result*
100	Médecin, Étude, Santé	*Doctor, Study, Health*
101	Droit, Gouvernement, Ministre	*Law, Government, Minister*
102	Ordinateur, Informatique, Sécurité	*Computer, Computer, Security*
103	Menu, Carte, Table	*Menu, Menu, Table*
104	Étudier, Professeur, Programme	*Study, Professor, Program*
105	Foi, Dossier, Recherche	*Faith, File, Research*
106	Gare, SNCB, Parking	*Train station, SNCB, Parking*
107	Victime, Plainte, Médecin	*Victim, Complaint, Doctor*
108	Numérique, n/a, n/a	*Numerical, n/a, n/a*

CLASSE	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	MOTS THÉMATIQUES CANDIDATS (ANGLAIS)
109	Scène, Guerre, Passage	*Scene, War, Passage*
110	Magistrat, Police, Justice	*Magistrate, Police, Justice*
111	Clijsters, Masters, Henin	*Clijsters, Masters, Henin*
112	Conseil, Gouvernement, Ministre	*Council, Government, Minister*
113	Scientifique, Médical, Famille	*Scientific, Medical, Family*
114	Système, Informatique, Client	*System, Computer, Client*
115	Droit, Foi, Pouvoir	*Law, Faith, Power*
116	Musulman, Religion, Islam	*Muslim, Religion, Islam*
117	Arabe, Ligue, Islamiste	*Arabic, League, Islamic*
118	Médecine, Recherche, Médical	*Medicine, Research, Medical*

Tableau 3.6. Traduction des termes thématiques extraits.

3.3.2.1. Présentation de la mesure d'évaluation

Plusieurs mesures fondées sur la base de données lexicale WORDNET permettent d'évaluer la proximité sémantique entre des couples de termes. Parmi les mesures les plus fréquemment citées, on note celles de Banerjee et Pedersen (2002), de Hirst et St-Onge (1998), de Leacock et Chodorow (1998), de Lin (1998), de Jiang et Conrath (1997), de Resnik (1995) et Wu & Palmer (1994)[7]. Toutes ces mesures permettent de calculer la proximité sémantique entre des couples de termes; elles se distinguent cependant par les méthodes sur lesquelles elles reposent. Dans tous les cas, la proximité entre les termes est représentée par une valeur numérique.

Afin d'évaluer la proximité sémantique entre les termes des catégories attribuées manuellement à chaque classe et les termes thématiques candidats extraits de chacune des classes, nous avons retenu uniquement la mesure développée par Hirst et St-Onge (1998). Pour réaliser ce processus d'évaluation, nous avons utilisé la version de cette mesure implémentée par Ted Pederson[8]. Le choix de cette mesure est motivé par deux principales raisons. La première réside dans le fait qu'elle permet d'obtenir une valeur normalisée. En effet, selon cette mesure, la proximité entre un couple de termes est reflétée par une valeur se situant entre 0 et 16. La valeur représentant le degré de proximité est donc très facilement

[7] Pour une description détaillée des principales caractéristiques de ces mesures, nous référons le lecteur au mémoire de maîtrise de Patwardhan (2003).

[8] wn-similarity.sourceforge.net

interprétable (ce qui n'est pas le cas de la majorité des autres mesures fréquemment citées). La seconde raison, plus importante encore, réside dans le fait que cette mesure ne permet d'identifier que des relations de proximité sémantique forte entre des termes. Ainsi, selon cette mesure la valeur 0 témoigne de l'absence de relation sémantique entre deux termes. Les autres valeurs possibles (comprises entre 1 et 16, inclusivement) témoignent toutes d'une proximité sémantique forte et sont associées à trois principaux degrés de proximité. Ainsi, selon cette mesure, deux termes peuvent 1) n'entretenir aucune relation de proximité sémantique, 2) être très fortement reliés (*extra-strong*), 3) être fortement reliés (*strong*) ou 4) être moyennement-fortement reliés (*medium-strong*).

Le calcul de proximité est réalisé en tenant compte de trois catégories de relations possibles entre les *synsets*. Il s'agit des relations ascendantes (*upward*), descendantes (downward) et horizontales (*horizontal*). Le tableau 3.7 présente les différentes relations possibles selon ces trois catégories directionnelles.

RELATION	CATÉGORIE DIRECTIONNELLE
Also see	Relation horizontale
Antonymy	Relation horizontale
Attribute	Relation horizontale
Cause	Relation descendante
Entailment	Relation descendante
Holonymy	Relation descendante
Hypernymy	Relation ascendante
Hyponymy	Relation descendante
Meronymy	Relation ascendante
Pertinence	Relation horizontale
Similarity	Relation horizontale

Tableau 3.7. Les relations et les catégories directionnelles dans WORDNET (tiré de Hirst et St-Onge, 1998, p. 308).

À partir de ces informations, Hirst et St-Onge définissent les différents degrés de relation possible de la manière suivante :

1. Deux termes sont très fortement reliés (relation extra-forte (*extra-strong*)) lorsque qu'ils sont littéralement identiques (la valeur représentant la proximité entre deux termes sera alors de 16).

2. Deux termes sont fortement reliés (relation forte (*strong*)) dans trois contextes précis :

a) lorsque les deux termes partagent un *synset* commun (figure 3.9 (a)) (la valeur représentant la proximité entre deux termes sera alors de 16) ;

b) lorsque les *synsets* des deux termes sont reliés selon une relation horizontale (figure 3.9 (b)) (la valeur représentant la proximité entre deux termes variera en fonction de la nature de la relation unissant les deux termes) ;

c) lorsque qu'un terme est une composante d'un terme complexe (figure 3.9 (c)) (la valeur représentant la proximité entre deux termes sera alors de 16).

3. Deux termes sont très moyennement-fortement reliés (relation moyenne-forte (*medium-strong*)) lorsqu'au moins un des *synsets* associés à chaque terme peuvent être reliés selon un chemin (*path*) précis. Les chemins admissibles sont présentés dans la figure 3.10[9]. Hirst et St-Onge justifient la validité de ces chemins ainsi (1998, p. 309) :

> *If a multilink path between two synsets is to be indicative of some reasonable semantic proximity, the semantics of each lexical relation must be taken into consideration. Now, an upward direction corresponds to generalization. For instance, an upward link from {*apple*} to {*fruit*} means that {*fruit*} is a semantically more general synset than {*apple*}. Similarly, a downward link corresponds to specialization. Horizontal links are less frequent than upward and downward links; a synset rarely has more than one. Such links are usually very specific of meaning. [...] So, to ensure that a path corresponds to a reasonable relation between the source and the target word, two rules have been defined as to which patterns are allowable:*
>
> *(R1) No other direction may precede an upward link.*
>
> *Once a link that narrows down the context (downward or horizontal) has been used, it is not permitted to enlarge the context again by using an upward link.*
>
> *(R2) At most one change of direction is allowed.*
>
> *Changes of direction constitute large semantic steps. Therefore, they must be limited. However, this second rule has the following exception:*
>
> *(R2') It is permitted to use a horizontal link to make a transition from an upward to a downward direction.*
>
> *Horizontal links correspond to small semantic distance for words such as* height *and* high, *which are linked by an attribute relation. In this case, this exception to (R2) enables*

---

[9] Les auteurs dressent aussi une liste des chemins non-admissibles. Il n'est pas nécessaire pour bien comprendre le principe de leur mesure que nous les présentions ici.

connections between subordinates of height *and subordinates of* high. *Thus, it has been assumed that enabling such a connection between two superordinates does not imply too large a semantic step.*

Un exemple de relation moyennement-forte est présenté dans la figure 3.11. Contrairement aux deux autres catégories de relations, la valeur des relations moyennement-fortes peut être calculée selon la formule suivante (dans laquelle C et k sont des constantes) :

$$Poids = C - longueur\ du\ chemin - k \cdot nombre\ de\ changements\ de\ direction \quad (6)$$

Selon ce calcul, plus un chemin entre deux termes sera long ou plus il comportera de changements de direction, plus le poids représentant la proximité sémantique entre les termes sera faible.

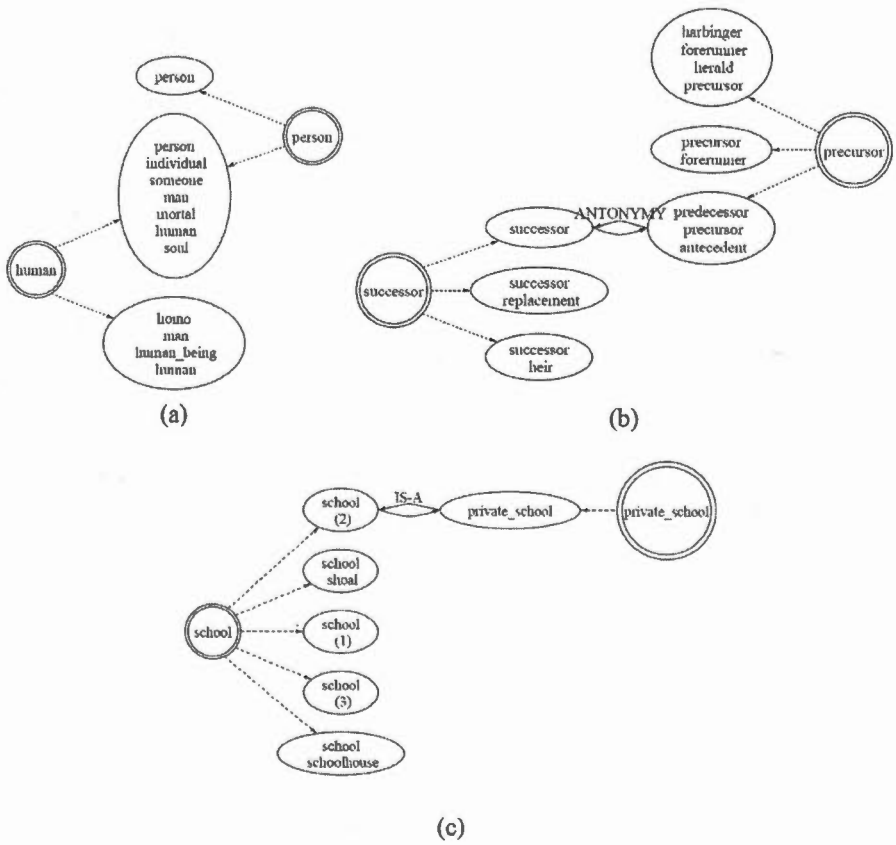


Figure 3.9. Exemple de trois relations fortes (tiré de Hirst et St-Onge, 1998, p. 310-311).

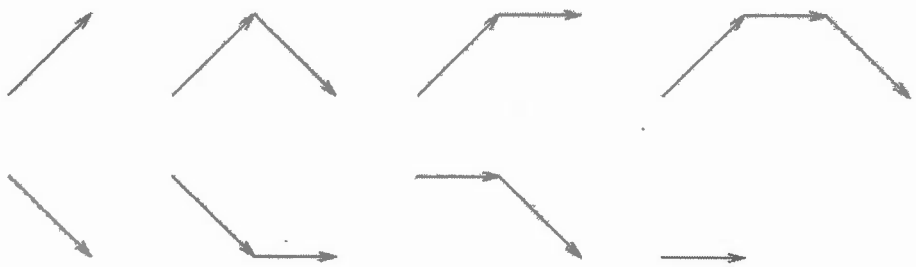


Figure 3.10. Les chemins admissibles pour les relations moyennement-fortes (tiré de Hirst et St-Onge, 1998, p. 312).

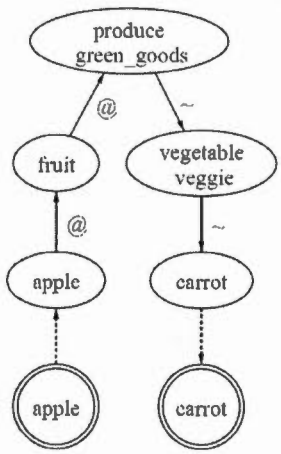


Figure 3.11. Exemple de relation moyennement-forte (@ = hyperonymie et ~ = hyponymie) (tiré de Hirst et St-Onge, 1998, p. 313).

3.3.3. Discussion des résultats de l'expérimentation 2

Les résultats que nous avons obtenus en employant la méthodologie fondée sur la classification automatique et sur l'identification automatique des termes thématiques candidats sont des plus concluants. En effet, selon cette mesure, et si l'on considère les termes extraits ne figurant pas dans la base de données WORDNET comme étant de mauvais termes thématiques, nous retrouvons seulement 30 classes pour lesquelles aucun des trois termes extraits n'est sémantiquement relié aux termes représentant les catégories attribuées manuellement. On retrouve donc 88 classes pour lesquelles il y a au moins 1 terme extrait (sur une possibilité de 3) qui est sémantiquement relié aux termes représentant les catégories attribuées manuellement. Plus spécifiquement, on retrouve 42 classes pour lesquelles un des

trois termes extraits est sémantiquement relié aux termes représentant les catégories attribuées manuellement; 33 classes pour lesquelles deux des trois termes extraits sont sémantiquement reliés aux termes représentant les catégories attribuées manuellement et 13 classes dont les trois termes extraits sont sémantiquement reliés aux termes représentant les catégories attribuées manuellement.

Nous avons de plus procédé à une évaluation manuelle des quelques termes candidats ne figurant pas dans la base de données lexicale WORDNET. Nous avons alors évalué subjectivement la présence de relations sémantiques entre les termes thématiques candidats absents de WORDNET et les catégories attribuées manuellement. Lorsque l'on considère les termes ne figurant pas dans la base de données lexicale WORDNET, les résultats sont encore plus encourageants. Ainsi, en tenant compte de ces mots, on note seulement 24 classes pour lesquelles aucun des trois termes extraits n'est sémantiquement relié aux termes représentant les catégories attribuées manuellement et donc 94 classes pour lesquelles il y a au moins 1 termes (sur une possibilité de 3) extrait qui est sémantiquement relié aux termes représentant les catégories attribuées manuellement. Plus spécifiquement, on retrouve 32 classes pour lesquelles un des trois termes extraits est sémantiquement relié aux termes représentant les catégories attribuées manuellement; 46 classes pour lesquelles deux des trois termes extraits sont sémantiquement reliés aux termes représentant les catégories attribuées manuellement et 16 classes dont les trois termes extraits sont sémantiquement reliés aux termes représentant les catégories attribuées manuellement.

CLASSE	CATÉGORIE ATTRIBUÉE MANUELLEMENT	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	VALIDITÉ	EXPLICATION
001	Informatique	Microsoft	Oui	Compagnie œuvrant dans le secteur de l'informatique
007	Informatique	Microsoft	Oui	Compagnie œuvrant dans le secteur de l'informatique
012	Pédophilie	Dutroux	Oui	Principal pédophile impliqué dans l'affaire Dutroux
		Nihoul	Oui	Inculpé dans l'affaire Dutroux
014	Électricité	Électrabel	Oui	Compagnie productrice et distributrice d'électricité
016	Tennis	Clijsters	Oui	Joueuse de tennis
		Henin	Oui	Joueuse de tennis
021	Tennis	Clijsters	Oui	Joueuse de tennis
027	Tennis	Henin	Oui	Joueuse de tennis
		Clijsters	Oui	Joueuse de tennis
		Williams	Oui	Joueuse de tennis
028	Tennis	Williams	Oui	Joueuse de tennis
029	Médecin	UCL	Non	Terme rejeté
030	Électricité	Électrabel	Oui	Compagnie productrice et distributrice d'électricité
033	Tennis	Clijsters	Oui	Joueuse de tennis
039	Pédophilie	Bourlet	Oui	Procureur du Roi dans l'affaire Dutroux
050	Université	ULB	Oui	Université Libre de Bruxelles
051	Gare	SNCB	Oui	Société Nationale des Chemins de fer de Belgique
054	Islam	AKP	Oui	Parti pour la justice et le développement (Parti politique turc de centre-droit apparemment associé au mouvement islamiste)
069	Gare	SNCB	Oui	Société Nationale des Chemins de fer de Belgique
075	Tennis	Henin	Oui	Joueuse de tennis
094	Médecin	Bourgmestre	Non	Terme rejeté
106	Gare	SNCB	Oui	Société Nationale des Chemins de fer de Belgique
111	Tennis	Clijsters	Oui	Joueuse de tennis
		Henin	Oui	Joueuse de tennis

Tableau 3.8. Résultats de l'évaluation subjective des termes thématiques absents de WORDNET.

Ces informations sont représentées graphiquement dans la figure 3.12 (voir l'annexe 6 pour les résultats détaillés de chaque classe individuelle).

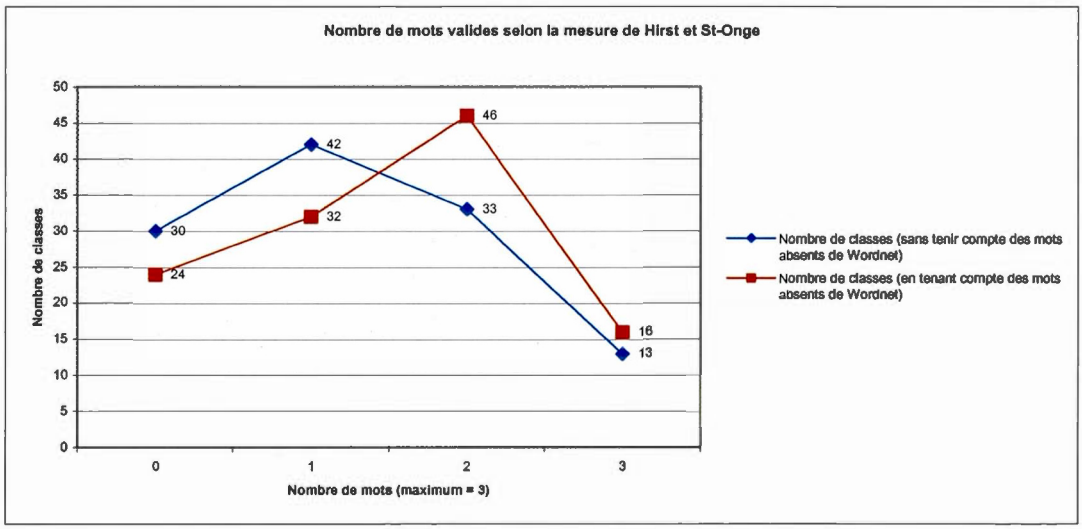


Figure 3.12. Nombre de mots valides selon la mesure de Hirst et St-Onge.

Les figures suivantes (figures 3.13 et 3.14) représentent graphiquement les résultats pour chacune des 10 catégories thématiques présentes dans le corpus d'expérimentation. La première de ces figures ne tient pas compte des mots absents de la base de données lexicales WORDNET. Cette figure nous indique clairement que la majorité des catégories (8 catégories sur 10) sont caractérisées par des résultats globalement comparables. Toutefois, deux catégories se démarquent de l'ensemble. La première, la catégorie « PÉDOPHILIE », en couleur bourgogne, présente des résultats très décevants. Comme nous pouvons le constater dans cette figure, aucun terme thématique candidat n'est sémantiquement relié au terme employé lors de la catégorisation manuelle. Nous pouvons expliquer ces résultats principalement par le fait que le thème de la pédophilie n'est pas, à tout le moins dans le corpus étudié, caractérisé par un vocabulaire précis et exclusif. En effet, les principaux termes qui furent extrait des documents traitant de la pédophilie sont les suivants : « Autorité », « Avocat », « Bourlet », « Chef », « Client », « Collège », « Dénoncer », « Directeur », « Dossier », « Dutroux », « Enfant », « Enseignement », « Judiciaire », « Juge », « Loi », « Nihoul », « Plainte », « Procès », « Procureur », « Professionnel », « Projet », « Tribunal »

et « Victime ». On constate que la majorité des termes extraits relèvent plutôt du domaine juridique que de la problématique de la pédophilie. Lorsque nous considérons le genre des textes analysés (et surtout le contexte dans lesquels ces articles journalistiques furent rédigés (couverture médiatique de « l'affaire Dutroux »)), l'emploi de ces termes associés au domaine légal n'est pas surprenant. Cependant, malgré la proximité que nous qualifions de « pragmatique » entre les termes extraits et la véritable catégorie des documents qu'ils composent, il n'en demeure pas moins qu'au niveau strictement sémantique il n'y a pas de relation satisfaisante entre des termes tels que « Procureur » ou « Procès » et l'étiquette thématique « PÉDOPHILIE ». Au niveau du processus évaluatif, la méthode employée ne permet pas d'attester d'un lien sémantique entre les termes automatiquement extraits et l'étiquette thématique attribuée manuellement. Cela ne signifie pas pour autant que les termes extraits ne représentent pas adéquatement le contenu des documents auxquels ils sont associés.

La seconde catégorie se distinguant des autres est la catégorie « INFORMATIQUE » (en couleur orange). Il s'agit de la catégorie pour laquelle un maximum de termes thématiques valides furent automatiquement extraits. Dans ce cas, nous pouvons expliquer ces résultats principalement par le fait que le thème de l'informatique est caractérisé par un vocabulaire précis et très exclusif. En effet, les principaux termes qui furent extrait des documents traitant de l'informatique sont les suivants : « Informatique », « Internet », « Logiciel », « Microsoft », « Numérique », « Ordinateur », « Réseau », « Sécurité », « Société », « Système », « Technologie », « Texte » et « Virus ». Il est manifeste que la majorité des termes extraits concernent directement la problématique de l'informatique.

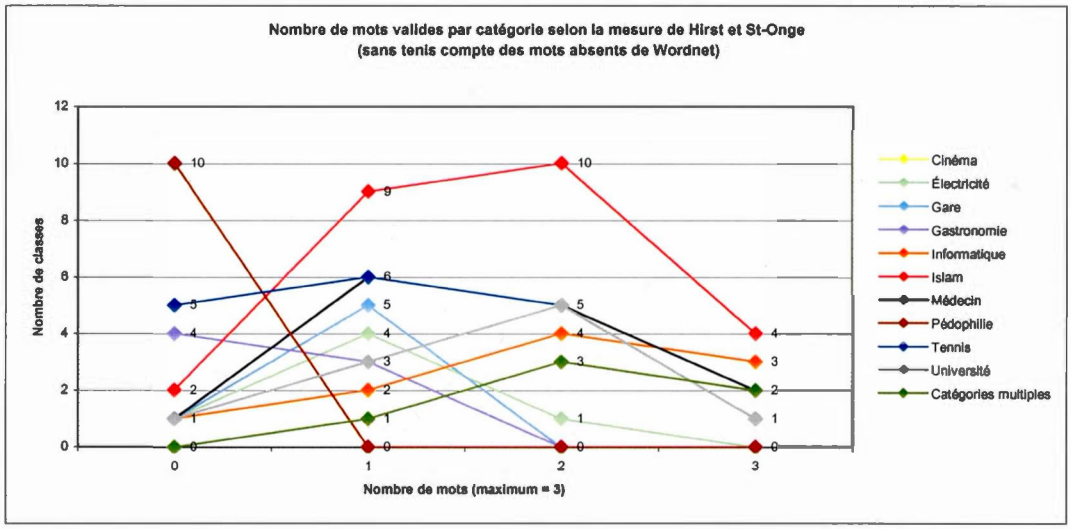


Figure 3.13. Nombre de mots valides par catégorie selon la mesure de Hirst et St-Onge (sans tenir compte des mots absents de WORDNET).

L'analyse des résultats obtenus en tenant compte des mots absents de WORDNET, nous indique aussi que la plupart des catégories sont caractérisées par des résultats globalement comparables. De manière générale, les résultats sont légèrement supérieurs. En ce qui concerne la catégorie « PÉDOPHILIE », les résultats ne sont cependant guère meilleurs. Seulement deux classes associées à cette catégorie ont été caractérisées par des termes valides. Les résultats de la catégorie « INFORMATIQUE » sont légèrement supérieurs (principalement en raison du mot « Microsoft »). Finalement, il importe de mentionner que la catégorie pour laquelle on constate la plus grande amélioration en tenant compte des mots absents de la base de données WORDNET est la catégorie « TENNIS ». Ainsi, grâce à cette stratégie, le nombre de classes associées manuellement à cette catégorie pour lesquelles aucun terme extrait n'a été jugé valide est passé de 5 à 2.

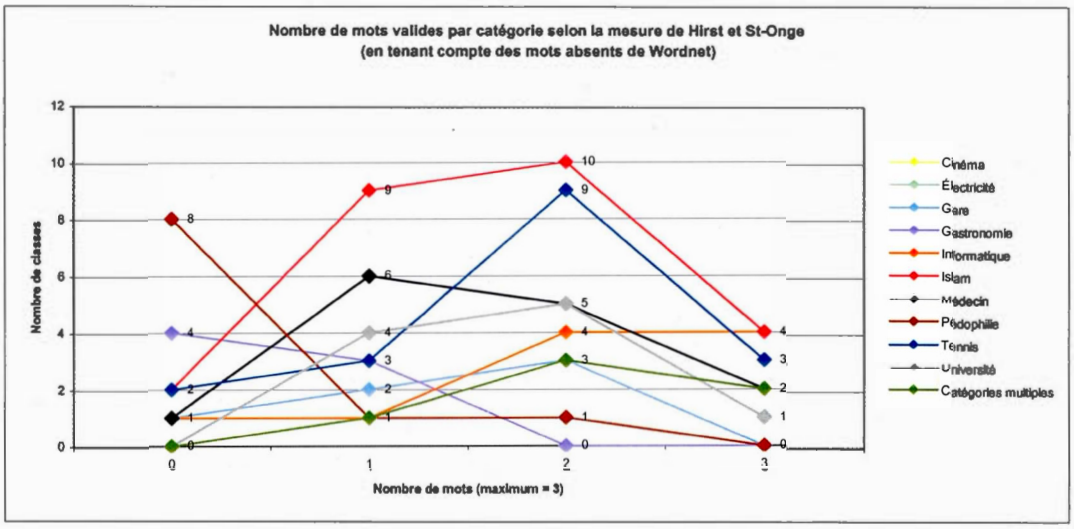


Figure 3.14. Nombre de mots valides par catégorie selon la mesure de Hirst et St-Onge (en tenant compte des mots absents de WORDNET).

CHAPITRE 4

APPLICATION DES RÉSULTATS À DES FINS D'IDENTIFICATION DE THÈMES ET D'ANALYSE THÉMATIQUE

Dans le deuxième chapitre, nous avons exposé deux méthodologies permettant de regrouper des segments de documents et d'identifier automatiquement le contenu thématique des différents regroupements. Nous rappelons que la première démarche méthodologique a été réalisée en employant une opération de catégorisation automatique. L'objectif principal de cette opération est, comme nous l'avons vu, d'attribuer automatiquement, à partir de métadonnées connues de certains documents, des étiquettes thématiques aux différents intrants soumis au système. En effectuant une telle opération, le processus de catégorisation permet donc de regrouper les intrants sur la base des différentes catégories possibles. La seconde démarche, fondée sur la classification automatique de segments documents et sur l'extraction automatique des termes thématiques candidats, permet d'obtenir des résultats de nature comparable mais, pour atteindre cet objectif, elle procède selon une autre stratégie. Ainsi, la seconde démarche vise d'abord à regrouper les segments de documents, puis elle permet, grâce à l'opération d'extraction automatique des termes thématiques candidats, d'identifier le contenu thématique des différents regroupements.

Dans le troisième chapitre, nous avons appliqué ces deux méthodologies sur un corpus d'articles de journaux. Nous avons alors observé une différence significative entre les résultats obtenus en appliquant les deux démarches méthodologiques. Nous avons observé que la démarche fondée sur les opérations de classification et d'extraction automatique des termes thématiques candidats, tout en ne faisant intervenir aucune information connue *a priori* (métadonnées) sur le contenu des documents, permet d'obtenir de meilleurs résultats que ceux obtenus par l'approche fondée sur l'opération de catégorisation automatique. Afin d'explorer la pertinence des opérations de regroupement et d'identification des thèmes à des

fins d'analyse thématique de documents textuels, nous avons retenu dans ce chapitre uniquement les résultats obtenus par la seconde démarche méthodologique.

Afin de représenter graphiquement les résultats obtenus lors des opérations précédentes et de démontrer la pertinence de ces opérations à des fins d'analyse thématique de documents textuels, nous avons employé une application de représentation graphique utilisant certains concepts et principes provenant de travaux sur les réseaux sociaux. L'application employée, nommée UCINET, peut être gratuitement téléchargée sur Internet[1] (comme la majeure partie des applications utilisées dans notre projet). Il est important de noter que nous avons employé uniquement une composante spécifique de cette application (composante nommée NETDRAW) conçue à des fins de représentation graphique d'information. Dans le cadre de ce projet, nous ne nous sommes pas attardé sur les fondements théoriques des travaux sur les réseaux sociaux. Nous référons les lecteurs intéressés à connaître davantage ce domaine de recherche aux travaux de Scott (2000), de Degenne et Forsé (1994) et de Wasserman et Faust (1995). Notons, par ailleurs, que quelques études ont déjà démontré la pertinence des techniques d'analyse et de représentation graphique reposant sur les théories des réseaux sociaux dans leur application à l'analyse des documents textuels (Archambeault, 2002; Popping, 2000).

Afin de pouvoir représenter graphiquement les résultats des opérations de classification et d'extraction automatique de termes thématiques candidats à des fins d'analyse thématique, nous avons dû procéder à une opération supplémentaire permettant de traduire les résultats de ces opérations dans un format compatible avec l'application UCINET. En effet, l'application UCINET n'accepte en intrant que des matrices de données. Nous avons donc traduit chaque classe de segments sous forme de vecteurs. Ainsi, nous avons à nouveau procédé au filtrage du lexique de chaque classe de segments selon les mêmes paramètres employés lors des opérations de filtrage réalisées antérieurement et nous avons opté pour une pondération des termes dans chacune des classes de segments en fonction de leur fréquence d'apparition dans ces dernières. Cette opération préliminaire nous a permis d'obtenir une matrice pondérée composée de 118 classes et de 143 termes thématiques (118 x 143). La dimension de la matrice soumise à l'application UCINET est supérieure aux autres matrices que nous avons

---

[1] www.analytictech.com

générées lors des processus antérieurs. Cette différence est motivée par les raisons justifiants les processus de filtrage des données. En d'autres termes, dans notre projet, nous avons réalisé deux opérations de filtrage visant à atteindre des objectifs radicalement différents. Afin d'effectuer les opérations de catégorisation et de classification, nous avons effectué un filtrage du lexique des segments dans le but de ne retenir que les termes les plus discriminants permettant de différencier et de regrouper les différents segments. L'objectif poursuivi était donc d'identifier, à partir de l'ensemble des termes de notre corpus, que ceux caractérisant le plus possible le contenu de chaque segment. En contre partie, l'opération de filtrage du lexique des classes de documents n'est pas effectuée afin d'atteindre le même objectif. Dans le cadre d'une tâche d'analyse thématique, notre objectif consiste à identifier les termes thématiques les plus pertinents, indépendamment de leur valeur discriminante au sein du corpus (lequel est maintenant composé de classes de segments). Concrètement, cela signifie que tout terme figurant dans une grande majorité de segments doit être automatiquement rejeté lorsque le filtrage est effectué dans une perspective de catégorisation ou de classification automatique des documents (ou des segments de documents). Cependant, ces mêmes termes fréquents dans une grande majorité de classes de documents (ou de segments de documents) ne doivent pas nécessairement être rejetés lors qu'il s'agit d'identifier la structure thématique d'un corpus. Il est possible que certains de ces termes, bien que possédant une valeur discriminante très faible, témoignent de la présence de thèmes importants du corpus. Dans cette perspective, ils doivent donc être retenus à des fins d'analyse thématique.

4.1. L'analyse thématique des documents textuels

Comme nous l'avons démontré dans le deuxième chapitre, plusieurs chercheurs se sont attardés sur la question de la thématique. Les nombreux travaux traitant de cette question s'inscrivent dans les disciplines académiques très variées (linguistique, littérature, etc.). En contrepartie, à notre connaissance, peu de travaux ont tenté de définir ou d'identifier les traits caractéristiques de l'activité d'analyse thématique des documents textuels. Comme nous l'avons démontré, la définition du concept de « thème » soulève des enjeux théoriques majeurs. Il ne semble pas y avoir de consensus sur la nature de ce à quoi nous référons en

employant les mots « thème » ou « thématique » et ce, même lorsque nous limitons notre investigation à un seul domaine de recherche. Par conséquent, définir l'activité d'analyse thématique s'avère autant, sinon davantage, problématique. En plus d'hériter des problèmes théoriques soulevés par le concept de « thème », circonscrire l'opération d'analyse thématique implique des considérations théoriques et pratiques supplémentaires que très peu de théories, tant linguistiques que littéraires, ont osé aborder. Plusieurs études ont en partie évacué la question de l'analyse thématique en la réduisant à l'activité d'identification des thèmes. Il n'est pas erroné d'affirmer que l'analyse thématique des données textuelles consiste à identifier les différents thèmes présents dans un corpus, mais l'opération d'analyse ne sautait se limiter uniquement à l'identification des thèmes abordés. Ainsi, nous soutenons que l'opération d'analyse thématique des documents implique des opérations cognitivement complexes telles que l'identification de la structure ou le réseau thématique observable dans un ensemble de textes. Il s'agit donc, en se basant sur l'opération d'identification des thèmes, d'identifier comment et, surtout, grâce à quelles unités linguistiques les thèmes sont reliés les uns aux autres. Comme nous le verrons, l'application d'une opération de regroupement (qu'elle soit réalisée par la classification ou la catégorisation) permet d'identifier les structures thématiques caractérisant un corpus textuel. Cependant, une telle approche, lorsqu'elle est réalisée en employant une représentation vectorielle des documents ne tenant pas compte de la nature linguistique des relations entre les termes, ne permet pas d'identifier la nature des relations unissant certains thèmes. En somme, l'approche proposée permet, en fonction des indices sémiotiques présents dans le texte, d'attester de la présence de relations entre différents thèmes, sans pour autant expliciter la nature de ces relations. À cet égard, lors de travaux antérieurs portant sur l'analyse thématique de textes philosophiques (Forest, 2002), nous avons démontré qu'il existait dans le *Discours de la méthode* de Descartes une relation entre le thème de la « distinction entre le corps et l'âme » et celui de l'« existence de Dieu ». Cependant, il nous semble difficilement réalisable d'assister, à l'aide des méthodologies présentées dans notre projet, l'identification de la relation entre ces deux thèmes spécifiques. En effet, une telle opération interprétative fait nécessairement appel à des considérations difficilement modélisables informatiquement. Bien qu'il soit possible d'identifier dans un corpus textuel certains marqueurs pouvant guider le processus d'interprétation, l'identification de la relation qu'entretiennent ces deux thèmes dans la

philosophie cartésienne fait appel à des connaissances philosophiques difficilement modélisable dans un langage informatique.

Dans le cadre de notre projet, ces réflexions nous incitent donc à concevoir l'opération d'analyse thématique en tant que processus permettant de découvrir et de parcourir les différents thèmes présents dans un corpus textuel. Comme nous l'avons évoqué dans nos travaux antérieurs (Forest, 2002), cette démarche est fort complexe. Elle repose en dernière instance sur plusieurs choix, tant théoriques que pratiques. Mais, de manière générale, l'analyse thématique consistera en un parcours caractérisé par un compromis entre, d'une part, les attentes du lecteur et, d'autre part, les indices sémiotiques présents dans le texte. Comme le souligne (Bremond, 1985, p. 420) :

> Par quoi suis-je orienté dans la série de mes choix? On peut répondre : par le désir d'isoler la ou les bonnes formes du thème. Mais qu'est-ce qu'une bonne forme? [...] la bonne forme, c'est celle qui procure la satisfaction la plus grande à mon attente de lecteur [...].

Bien qu'il la développe dans une tout autre perspective, Rastier semble aussi défendre cette idée en mentionnant :

> [...] il est clair que tout lexème n'est pas un thème. Une analyse thématique qui en resterait au palier lexical compterait potentiellement autant de thèmes que de mots de la langue, sauf bien sûr à restreindre cet inventaire, comme le font les dictionnaires de thématique, de façon normative et non-critique. On objectera que les thèmes sont ordinairement dénombrés par un lexème; mais ce lexème est simplement une lexicalisation privilégiée du thème, et l'on pourrait fort bien rencontrer des thèmes sans lexicalisation privilégiée.
>
> Comme toutes les unités sémantiques, un thème est une construction, non une donnée; aussi la thématique dépend des conditions herméneutiques : l'interprétation des données textuelles se place dans un cercle méthodologique dépendant du cercle herméneutique. Tout choix de corpus, tout prélèvement dans un corpus, tout recueil de données reste tributaire d'un choix qu'il importe de rendre explicite. En d'autres termes, pour atteindre ses objectifs, la thématique doit guider l'analyse lexicale, puis interpréter ses résultats qui sans cela resteraient inutilisables pour une sémantique textuelle. L'analyse lexicale, dont la statistique est un auxiliaire, ne propose pas d'elle-même des indices à l'analyse thématique. » (2001, p. 191)

Pour d'autres (Prince, 1985), l'attente du lecteur prendra le nom de « réalité extra-textuelle ». D'ailleurs plusieurs théoriciens ont noté l'importance, dans l'activité de

thématisation et de découverte des contenus thématiques, de cette composante essentiellement subjective.

> La nature de ce que l'on pourrait appeler plus généralement l'étude thématique des textes est d'abord fonction de l'objectif visé. (Martin, 1995, p. 18)

Dans le cadre de notre recherche, cette composante subjective se manifestera dans l'intérêt du chercheur envers certains mots thématiques qu'il privilégiera et dans les objectifs qu'il désire atteindre lors de son analyse. Ainsi le chercheur pourra, par exemple, choisir d'explorer un ou plusieurs thèmes précis du corpus, et ce dans le but d'en démontrer l'organisation, la structure, etc. Peut-être voudra-t-il explorer l'ensemble des thèmes d'un corpus afin d'identifier certaines informations précises concernant une problématique.

> Thématiser un texte dépend donc non seulement du « texte même » mais aussi (et peut-être davantage) du thématiseur, du cadre adopté, des unités choisies, des opérations accomplies pour les harmoniser, des résumés et paraphrases effectués. (Prince, 1985, p. 432)

Mais, d'autre part, une seconde contrainte, plus objective cette fois, entre aussi en compte dans le cadre de la tâche d'analyse et de découverte. Cette contrainte repose sur le texte à analyser. Cette composante, intra-textuelle, limite nécessairement la liberté du lecteur, car elle guide inévitablement l'ensemble des analyses. En effet, malgré les intérêts et les raisons qui mènent le lecteur ou l'analyste vers la découverte d'un thème particulier plutôt que d'un autre, le lecteur ou l'analyste ne crée pas les thèmes dans le corpus qu'il analyse. C'est le texte qui expose, à l'aide des différents porteurs sémiotiques qu'il comporte, les thèmes sur lesquels le chercheur posera éventuellement son analyse.

Dans les faits, à partir des résultats obtenus par la classification et l'identification du contenu thématique des classes, l'opération d'analyse thématique se déroule selon les étapes suivantes. Dans un premier temps, l'utilisateur choisi en fonction de ses intérêts de recherche ou d'analyse une classe thématique particulière caractérisée par un terme spécifique. À cette étape, il est possible de consulter les différents segments constituant le regroupement thématique (Figure 4.1).

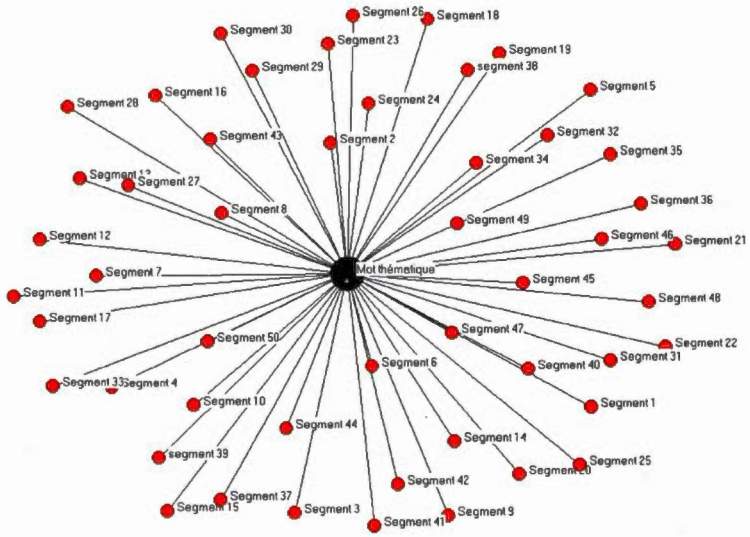


Figure 4.1. Choix d'un mot thématique et consultation des segments constituant la classe thématique.

Chaque classe est constituée d'un ou de plusieurs segments de textes qu'il est possible de consulter et desquels il est possible d'extraire le lexique thématiquement significatif (figure 4.2).

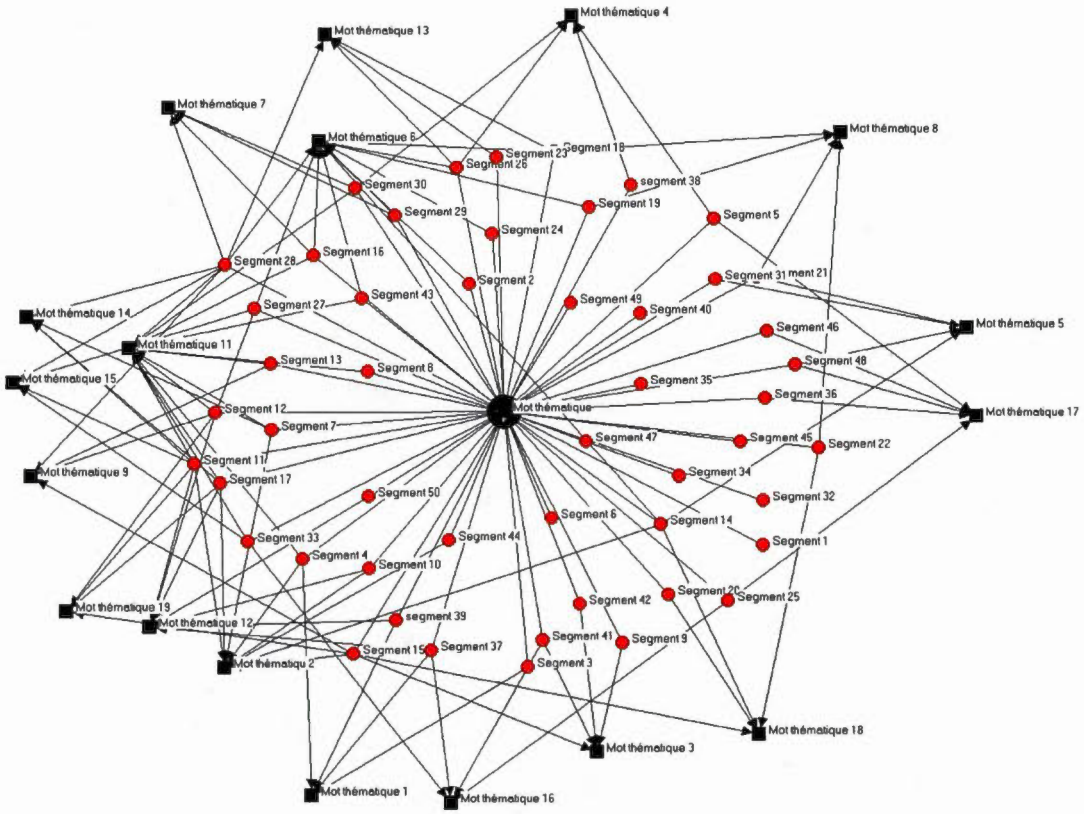


Figure 4.2. Choix d'un mot thématique, consultation des segments constituant la classe thématique et identification des mots thématiques de chaque segment.

On constate que les segments constituant un regroupement thématique sont composés, dans certains cas, de mots que l'on retrouve aussi dans d'autres segments présents dans d'autres regroupements thématiques. C'est à partir de ces mots apparaissant dans plus d'une classe thématique que la découverte des thèmes et de leur organisation est possible. Cela ouvre donc la voie à plusieurs cheminements thématiques possibles. En d'autres termes, certaines unités lexicales présentes dans une classe particulière peuvent se retrouver aussi dans une autre classe, indiquant par là qu'elles opèrent dans un autre contexte thématique.

4.2. Représentation des résultats globaux obtenus

Grâce à la démarche méthodologique et à l'application de représentations graphiques employées, il nous est possible d'obtenir une représentation générale de l'ensemble des parcours thématiques possibles dans notre corpus (figure 4.3).

Dans cette figure, chaque classe de segments est représentée par un point de couleur rouge. Comme nous l'avons mentionné, l'opération de classification a permis de regrouper l'ensemble des 1 625 segments en 118 classes. Les termes thématiques retenus suite au filtrage du lexique des classes sont représentés par des carrés de couleur noire. La disposition dans un plan en deux dimensions à été générée automatiquement par la fonctionnalité de représentation graphique de l'application UCINET. La taille des traits reliant les classes et les termes thématiques varie en fonction de la fréquence de chacun des mots dans chacune des classes.

À partir de cet aperçu général, nous pouvons consulter plus en détail la liste des termes retenus de chaque classe. Les figures 4.4 et 4.5 présentent graphiquement le lexique des classes 3 et 104 (les représentations graphiques des termes retenus dans chaque classe se trouvent dans l'annexe 7). Évidemment, à partir de ces représentations, il est toujours possible de consulter le texte de chaque segment (chaque segment étant identifié par un numéro unique). Dans les représentations détaillées de chaque classe, les cercles rouges représentent les segments de la classe, alors que les termes thématiques sont représentés par des carrés noirs.

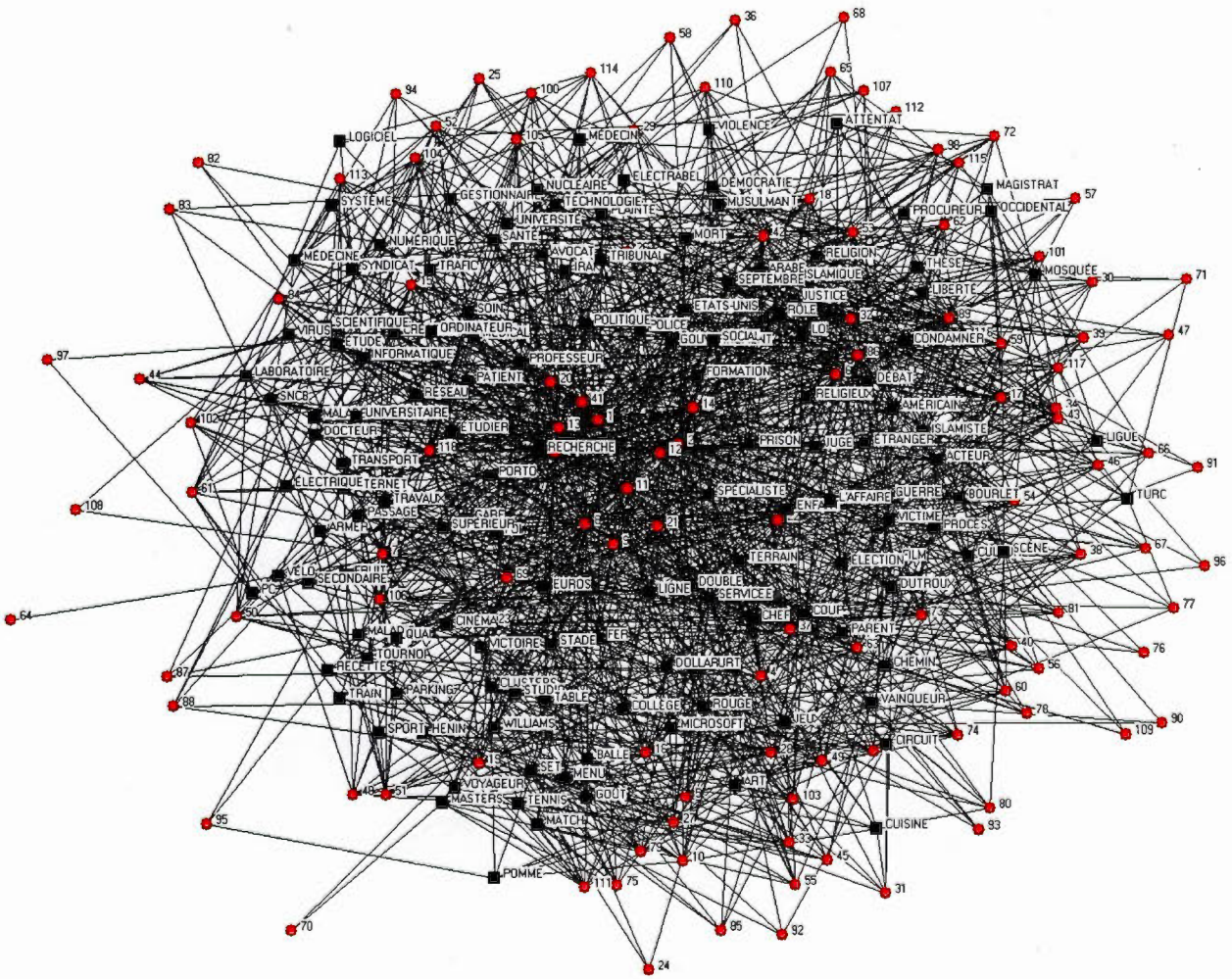


Figure 4.3. Représentation de l'ensemble des réseaux thématiques possibles dans le corpus (● = Classe et ■ = Mot).

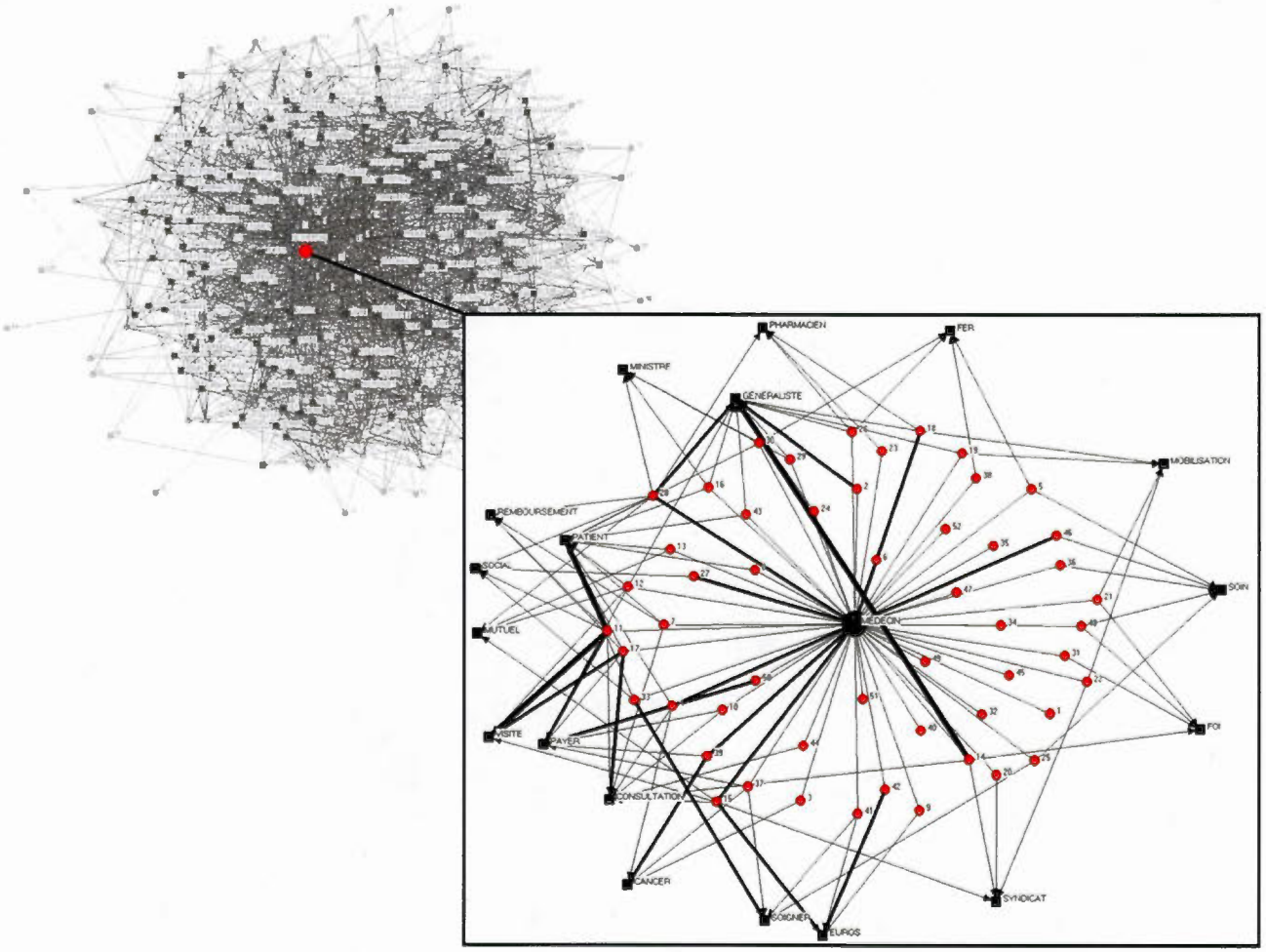


Figure 4.4. Représentation graphique du contenu de la classe 3 (● = Segment et ■ = Mot).

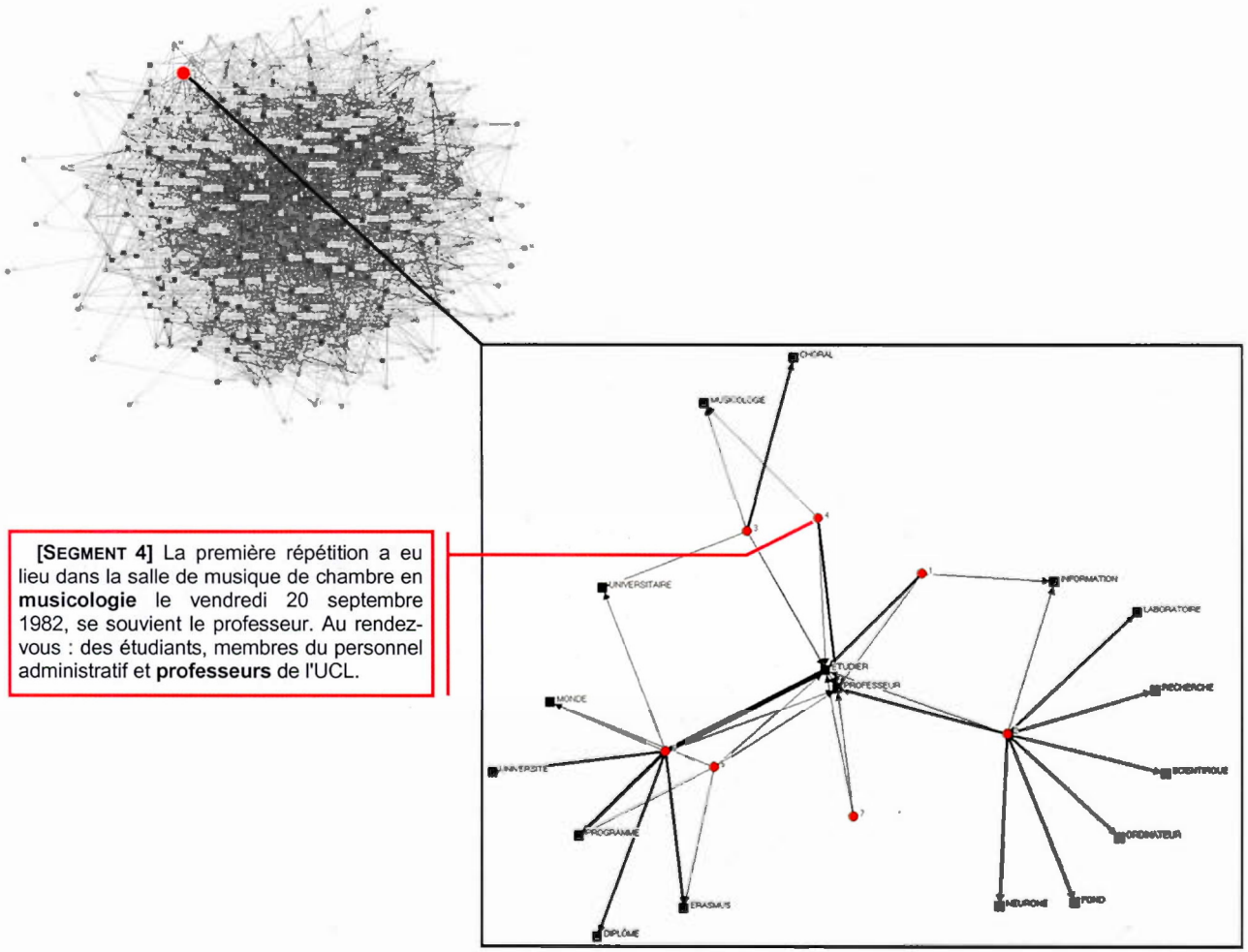


Figure 4.5. Représentation graphique du contenu de la classe 104 et consultation du segment 4 de cette classe (● = Segment et ■ = Mot).

Dans les pages suivantes, nous présentons les détails de quatre parcours thématiques différents qu'il est possible d'effectuer dans notre corpus.

4.3. Parcours thématique 1

Comme nous l'avons souligné, l'opération d'analyse thématique débute nécessairement par le choix d'un terme thématique particulier. Pour effectuer ce choix de départ, il peut être utile de consulter la carte thématique générale représentant tous les thèmes présents dans le corpus (figure 4.3) ou le tableau présentant la liste des termes thématiques de chaque classe (tableau 3.4, pp. 99-101). Dans notre corpus d'expérimentation, le choix du terme thématique « Électricité » représentant entre autres le contenu thématique de la classe 30 permet d'abord à l'utilisateur de consulter les différents documents abordant cette question.

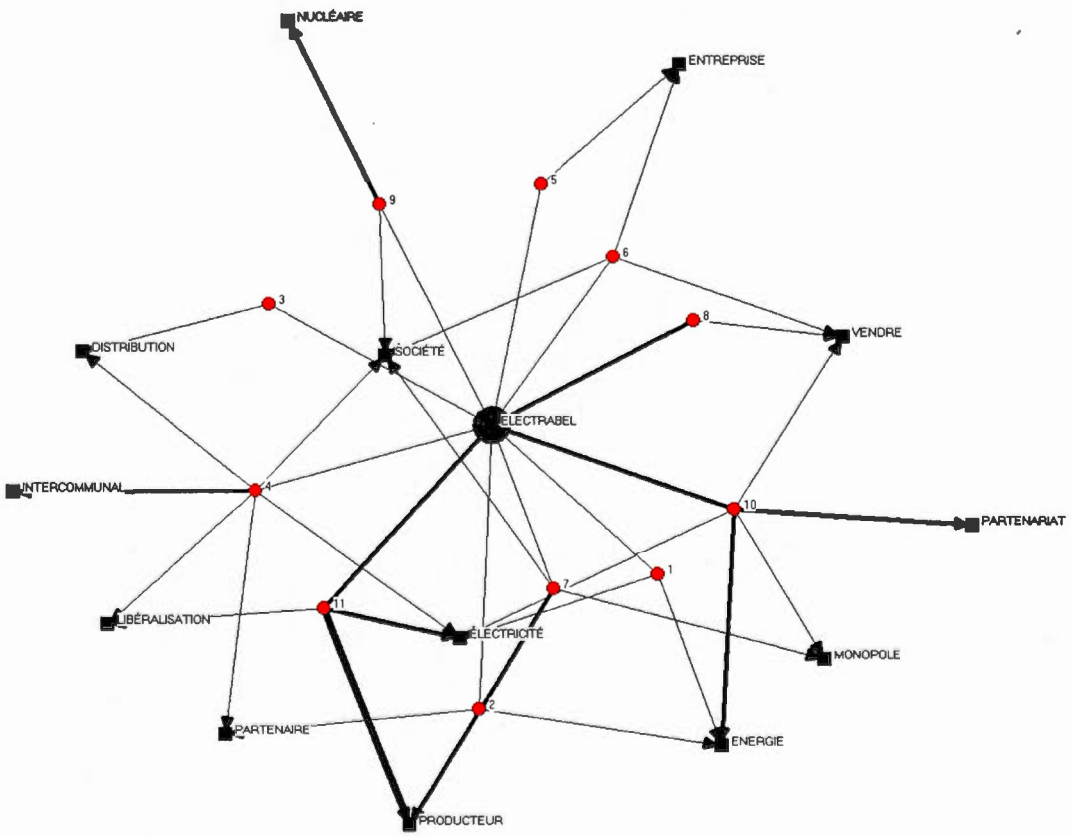


Figure 4.6. Représentation graphique du contenu thématique de la classe 30 (● = Segment et ■ = Mot).

Ainsi, en choisissant « Électricité » comme premier terme d'analyse, l'utilisateur peut consulter les segments de documents de la classe 30 tous reliés, de manière générale, au thème de l'électricité. Cependant, on constate qu'il est possible d'effectuer des sous-regroupements traitant de problématiques plus spécifiques, au sein d'une même classe.

Par exemple, on constate que le segment 9 de la classe 30 traite de l'électricité, mais, plus spécifiquement, de l'énergie nucléaire et des énergies non-polluantes.

> **[SEGMENT 9]** À Electrabel de rendre ses chiffres enfin publics. Pour faire avec exactitude ce type de comparaison en Belgique, il faudrait commencer par assainir le débat qui oppose actuellement partisans du nucléaire et défenseurs des énergies non polluantes. En particulier, mettre à la disposition de tous les données indispensables à ce débat de société. Or, l'industrie nucléaire s'est jusqu'ici retranchée dans une attitude de secret bien peu démocratique.

En contrepartie, on note que le segment 10 traite aussi du même sujet général, mais il aborde plus spécifiquement la question du partenariat dans la production et la distribution d'électricité.

> **[SEGMENT 10]** Cette opération n'est pas surprenante en soi. Electrabel avait, en effet, conclu en août 2001 un partenariat avec la CNR sous la forme d'une joint-venture commerciale, baptisée Energie du Rhône, et dans laquelle Electrabel avait 49% et CNR 51%. Energie du Rhône vend de l'électricité aux grands clients éligibles à l'ouverture du marché. Ce partenariat avait été créé dans le cadre de l'ouverture du marché français et de l'abolition du monopole de l'EDF.

Quant au segment 7, il traite plutôt de la question du monopole dans le domaine de l'électricité.

> **[SEGMENT 7]** Jusqu'en juin 2001, les poteaux et lignes qui transportaient les électrons appartenaient à la CPTE, une société détenue par les producteurs que sont Electrabel et SPE. Puis la CPTE s'est effacée pour faire place à Elia, une entité juridique autonome des producteurs mais toujours sous la coupe d'Electrabel. Ce qui ne manquait pas de susciter des frictions. Les concurrents d'Electrabel accusant Elia de défendre avant tout les intérêts d'Electrabel et de maintenir son quasi monopole.

Après avoir consulté le contenu des différents segments constituant la classe 30, l'utilisateur peut choisir un terme présent dans cette classe et explorer les autres classes pour

lesquelles ce terme à été retenu comme terme thématique. Ainsi, si l'utilisateur choisi le terme « Producteur », il sera dirigé vers la classe 31 dans laquelle le terme figure dans un tout autre contexte thématique (Figure 4.7).

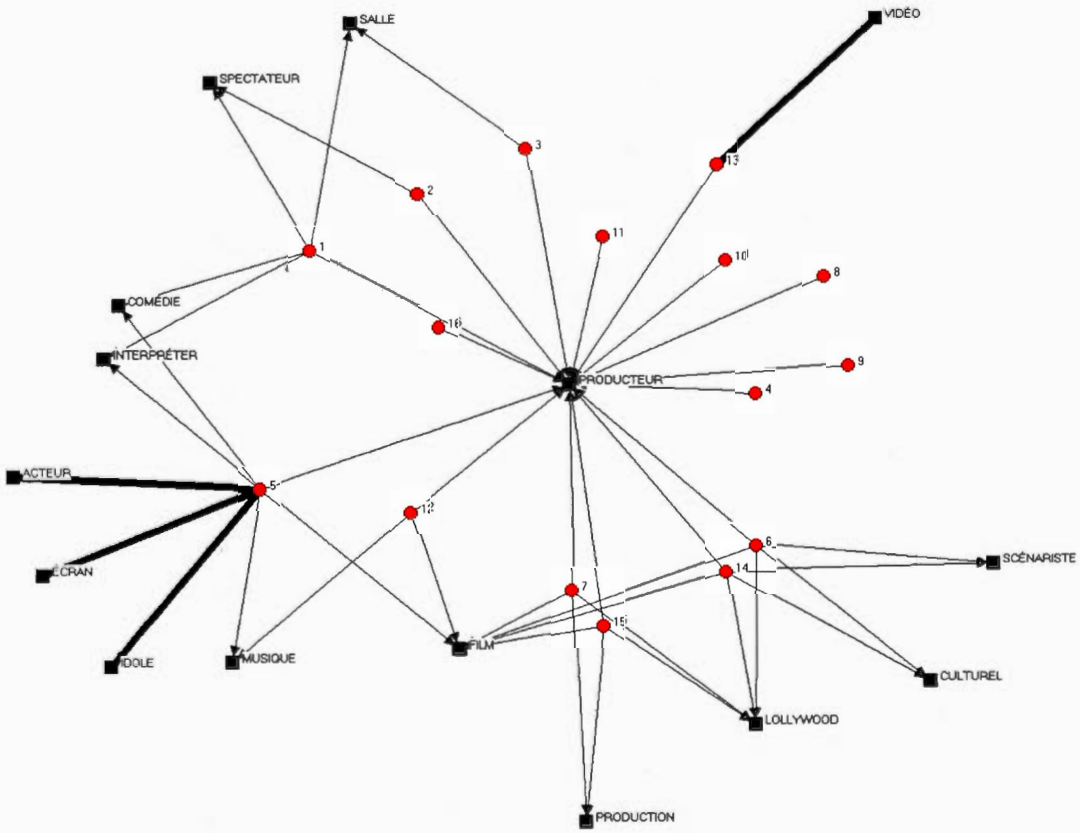


Figure 4.7. Représentation graphique du contenu thématique de la classe 31 (● := Segment et ■ = Mot).

Le terme thématique « Producteur » est présent dans deux classes thématiques distinctes traitant, dans la classe 30, de l'électricité et, dans la classe 31, de différents thèmes spécifiques qu'il est possible d'identifier en consultant les différents segments présents dans la classe. Il est donc présent dans au moins deux contextes thématiques différents, permettant ainsi à l'utilisateur de naviguer d'une classe à l'autre.

Comme c'était le cas dans la classe 30, il est à nouveau possible d'identifier dans la classe 31 des sous-regroupements correspondant à des contextes plus spécifiques dans lesquels le

terme « Producteur » est présent. À titre d'exemple, dans les segments 1, 2 et 3 de la classe 31, le terme « Producteur » est employé dans un contexte cinématographique.

> **[SEGMENT 1]** Acide et décalée, cette comédie déjantée (voir critique ci-dessous) est une réussite. Guillaume Canet invite le spectateur à partir à la rencontre d'un producteur de télévision et de reality show, Jean-Louis Broustal (interprété par un formidable François Berléand), aigri et manipulateur qui va inviter Bastien, le chauffeur de salle (c'est le personnage de Guillaume Canet), à passer un week-end, chez lui, à la campagne pour le manipuler comme c'est pas permis.

> **[SEGMENT 2]** En un long plan séquence, le spectateur se retrouve plongé sur le plateau d'Envoyez les mouchoirs, une émission de télé-réalité pour le moins controversée qui cartonne à l'audimat et chapeautée par son producteur, Jean-Louis Broustal (François Berléand).

> **[SEGMENT 3]** Bastien (Guillaume Canet) est le chauffeur de salle d'Envoyez vos mouchoirs et voue une admiration sans bornes à son producteur. Prêt à tout pour réussir, Bastien se laisse même humilier par Philippe Letzger (Philippe Lefebvre), l'animateur en question. Lorsque Broustal invite Bastien, chez lui, à la campagne pour un week-end de travail, ce dernier accepte sans hésiter au point de mettre sa vie affective en péril.

Au sein de la même classe, le terme « Producteur » guide aussi l'utilisateur vers d'autres contextes thématiques secondaires et moins caractéristiques du regroupement. Ainsi, il est présent dans des contextes écologique (segment 9) ou informatique (segment 13).

> **[SEGMENT 9]** Par contre, pour la Région qui a décidé de respecter le protocole de Kyoto réduction de 7,5% des gaz à effet de serre dont le CO2 les incinérateurs de déchets, producteurs de CO2, ne méritent pas d'obtenir un certificat vert.

> **[SEGMENT 13]** Le monde du jeu vidéo est en pleine effervescence. La preuve par l'exemple : le géant Microsoft, dont la stratégie commerciale repose de plus en plus sur le jeu vidéo, a acheté cette semaine au japonais Nintendo l'ensemble de ses parts dans le producteur anglais Rare Limited pour 375 millions de dollars. Et ce, alors que le numéro 1 français Infogrames, affaibli par un lourd endettement, pourrait envisager une importante augmentation de capital. A défaut, l'éditeur indépendant pourrait susciter bien des convoitises en forme d'OPA.

La découverte des thèmes et de leur organisation au sein du corpus peut aussi être réalisée en employant différents termes. Le choix du terme initial guide en partie l'analyse de l'utilisateur. Mais, comme nous le constatons, c'est l'utilisateur qui, en dernière instance, choisir les termes subséquents qui le dirigerons vers des univers thématiques *a priori*

imprévisibles. Dans notre premier exemple, le terme « Électricité » a permis de découvrir, sans grande surprise, le thème de l'électricité, de l'énergie nucléaire et des énergies non-polluantes. Mais, le choix du terme « Producteur » a mené l'utilisateur vers différents thèmes (non reliés à ceux de l'électricité et de l'énergie) dont ceux du cinéma (il s'agit alors du terme « Producteur » dans son acception cinématographique), de l'écologie (producteur de CO2) et de l'informatique (producteur de jeux vidéo).

4.4. Parcours thématique 2

L'emploi de termes thématiques différents guidera l'utilisateur vers d'autres univers thématiques du corpus. Ainsi, s'il choisit le terme « Tournoi » comme moyen d'accéder au contenu de la classe 16 (classe caractérisée par les termes thématiques « Tournoi », « Clijsters » et « Henin »), l'utilisateur pourra explorer le thème du tennis.

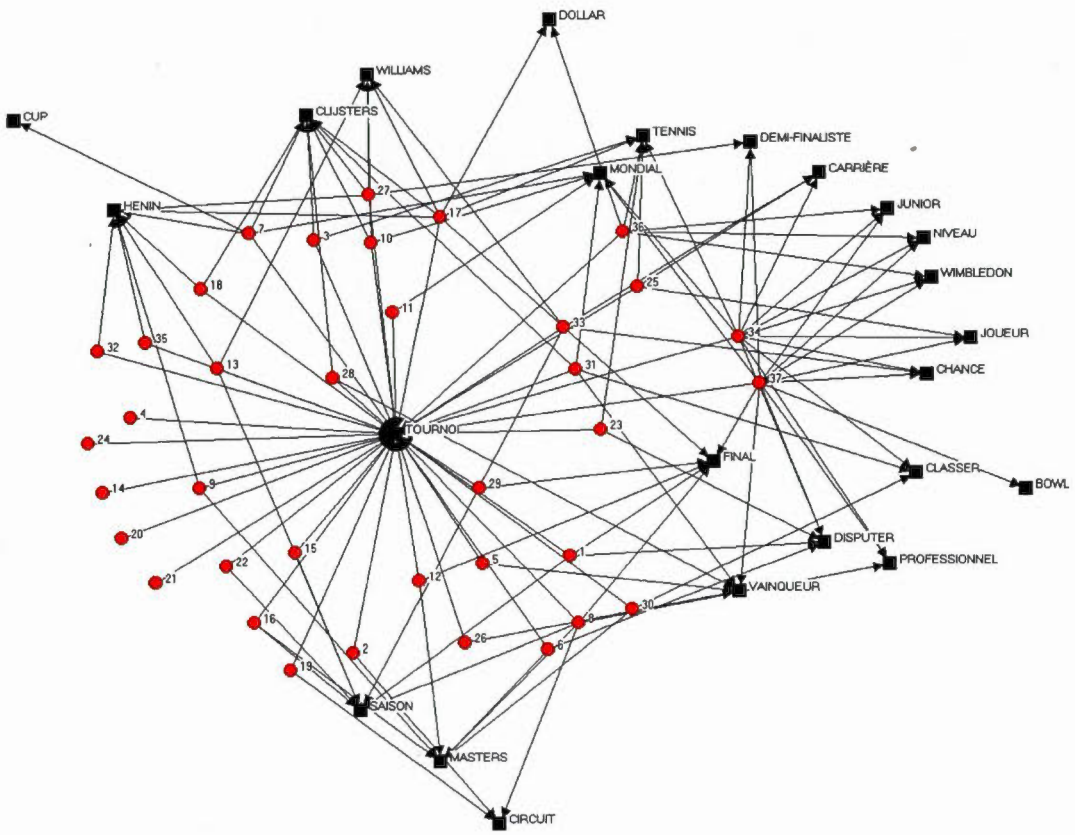


Figure 4.8. Représentation graphique du contenu thématique de la classe 16 (● = Segment et ■ = Mot).

Cette classe est composée de segments abordant tous le thème du tennis, comme en témoignent les passages suivants :

> **[SEGMENT 1]** Première joueuse à remporter trois finales consécutives à Melbourne (1997, 98, 99), Hingis y avait ensuite enchaîné trois finales supplémentaires, perdues celles-là. Lors de la dernière, disputée au mois de janvier dernier face à Jennifer Capriati, elle aurait même dû s'imposer si elle n'avait pas flanché sur la fin, victime, entre autres, des conditions dantesques (40° à l'ombre) dans lesquelles s'était déroulée cette rencontre. En grande forme, elle avait remporté, deux semaines auparavant, le tournoi de Sydney avant de remettre le couvert à Tokyo, quelques jours plus tard, face à Monica Seles. Ce qui laissait présager une agréable saison. Ce furent pourtant là ses deux seuls faits d'armes d'une année 2002 marquée par deux opérations aux ligaments de la cheville.

> **[SEGMENT 34]** Sous sa tignasse blonde comme les blés, Steve Darcis affiche, à 18 ans à peine, une belle assurance, que rien, en dehors des courts de tennis, ne semble réellement pouvoir ébranler. Posé et peu expansif, à défaut d'être replié sur lui-même, le jeune Liégeois sait qu'il est, aujourd'hui, au pied du filet. Les statistiques plaident en ma faveur. Elles indiquent que 70 % des joueurs classés dans le top 15 mondial chez les juniors parviennent à percer au niveau professionnel. J'ai donc toutes mes chances. Septième joueur mondial chez les juniors, le demi-finaliste du tournoi de Wimbledon, en juillet dernier, sera rapidement fixé sur son sort : c'est ce week-end qu'a démarré sa carrière pro dans un tournoi senior disputé en région parisienne.

Comme ce fut de cas dans le parcours thématique précédent, certains termes présents dans la classe 16 sont aussi présents dans d'autres classes (mais ce n'est pas nécessairement le cas – il est en effet possible d'atteindre une classe thématique ne permettant pas d'accéder à d'autres thèmes), indiquant par là qu'ils opèrent dans des contextes thématiques différents. Ceci est entre autres le cas du terme « Circuit », lequel, figurant aussi dans la classe 37, permet de découvrir le contenu thématique de cette nouvelle classe caractérisée par les termes thématiques « Gare », « Quartier » et « Voie ».

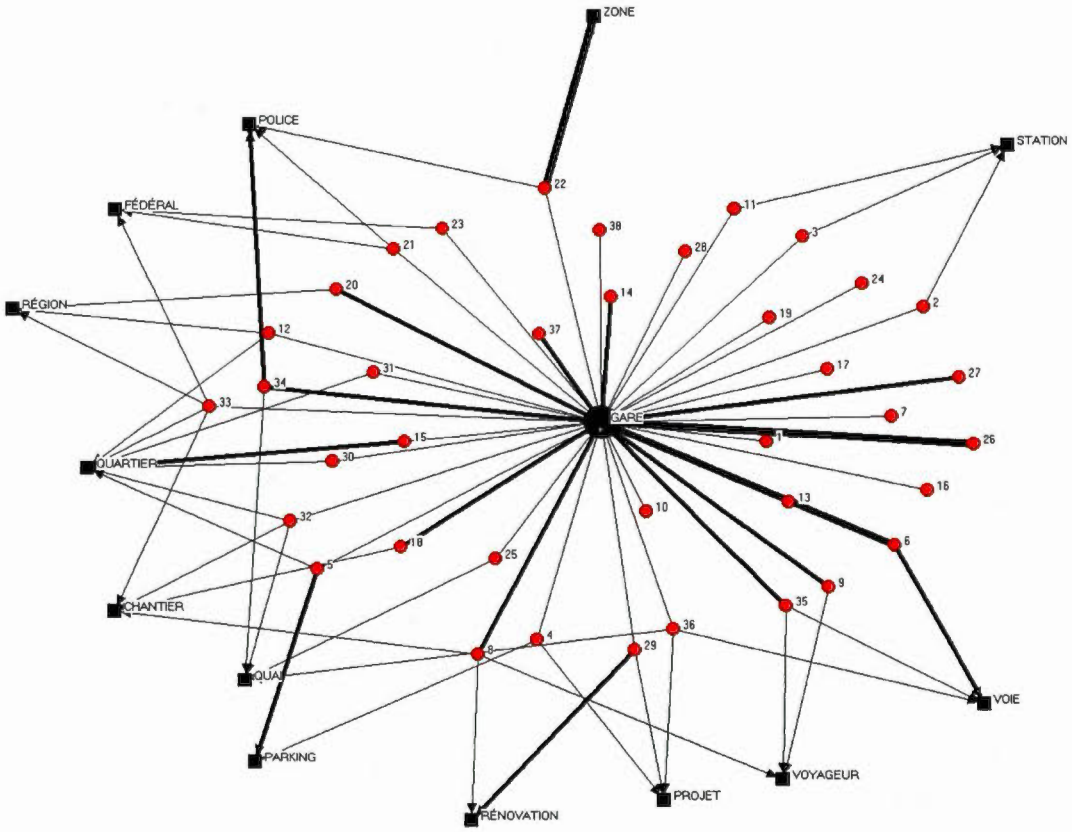


Figure 4.9. Représentation graphique du contenu thématique de la classe 37 (• = Segment et ■ = Mot).

Le segment 4 de la classe 37 est effectivement composé du terme « Circuit » dans un contexte associé cette fois aux domaines du transport et des gares.

[SEGMENT 4] Le projet prévoit même une extension des installations du TEC puisque le parking serait couvert d'une superstructure où serait aménagée une nouvelle gare des bus avec un circuit *kiss and ride* (un bisou et on s'en va).

[SEGMENT 8] Le vendredi 20 septembre prochain sera un jour béni entre tous pour les quelque 20 000 voyageurs qui s'embarquent quotidiennement à la gare de Namur. Il signifiera la fin de leur parcours du combattant. Celui auquel ils étaient soumis depuis des années par le vaste chantier de rénovation de la première gare wallonne. Dans sa dernière version car il y en a eu plusieurs depuis le démarrage des travaux en 1993 il consistait à gravir chaque matin un escalier métallique plutôt escarpé d'une quarantaine de marches.

Il est pertinent de noter que le terme « Circuit » est présent dans la figure 4.8, car, étant fortement présent dans cette classe, il a été retenu pour en représenter le contenu. Dans la classe 37 (figure 4.9), le terme est par contre peu fréquent. Il n'a donc pas été retenu pour en représenter le contenu. Bien qu'il soit absent de la figure 4.9, il n'en demeure pas moins que le terme « Circuit » est présent dans la classe 37. Il permet donc de naviguer de la classe 16 à la classe 37. La démarche que nous proposons est donc très flexible afin de permettre, d'une part, l'identification automatique du contenu des classes de documents et, d'autre part, la découverte et l'analyse thématique des différents thèmes, qu'ils soient très généraux ou plus spécifiques, du corpus soumis à l'analyse.

4.5. Parcours thématique 3

Dans notre corpus d'expérimentation, le choix du terme « Cuisine » représentant le contenu thématique de la classe 80, composée à 94% de segments de documents catégorisés manuellement par l'étiquette thématique « GASTRONOMIE », permet d'accéder à un autre univers thématique du corpus. En effet, en choisissant ce terme et cette classe thématique, il est possible de consulter les segments de documents dans lesquels on retrouve des mots tels que « Bouteille », « Chocolat », « Fromage », « Gastronomie », « Nourriture » et « Viande » (figure 4.10).

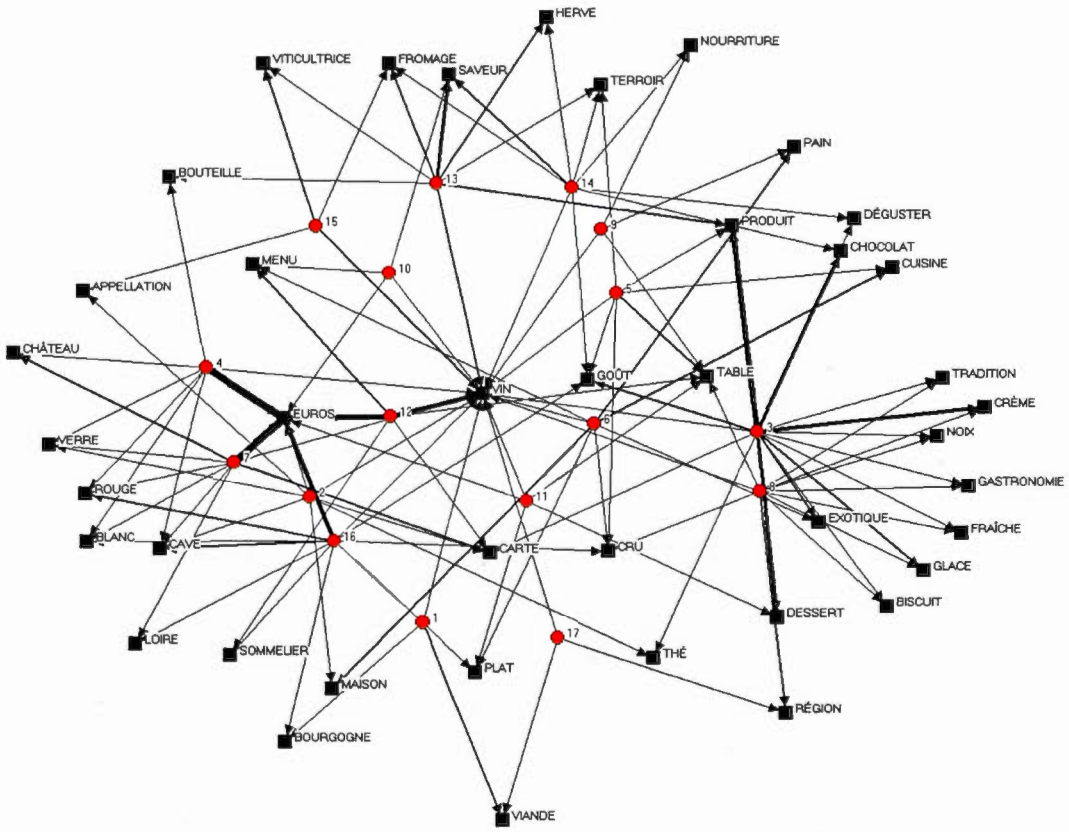


Figure 4.10. Représentation graphique du contenu thématique de la classe 5 (● = Segment et ■ = Mot).

Dans cette classe de segments de documents, il est possible de distinguer des sous-regroupements traitant de sujets plus spécifiques. Ainsi, il est possible de consulter des segments traitant des vins.

[SEGMENT 4] Nombreuses pages sous feuilles plastifiées, ce qui n'est jamais agréable. Le vin du mois est proposé au verre et à la bouteille. En blanc, un Côtes de Ventoux à 14,70 euros, le Saint-Véran 2000 de Drouhin à 29,50 euros, le sancerre 1999 Grande Réserve de Bourgeois à 36 euros. En rouge, on s'amuse des côtes roannaises 98, sélection de Pierre Troisgros à 22,40 euros. Moulis assez bien structuré, le Château Malmaison 1997 apporte les satisfactions vineuses espérées à 35,50 euros. Le Santenay 1990 de Nouveau s'impose s'il n'a pas trop évolué à 40,90 euros.

[SEGMENT 7] Des vins tricolores, qu'on commande au verre, au quart et en demi-bouteille, de 2,6 à 10 euros. Carte de type court, courant sur deux pages, où les idées ne manquent cependant pas. En blanc, le Michel Lynch défend les couleurs bordelaises à 24 euros, le pavillon bourguignon étant bien représenté par un Saint-Aubin 2000 du Domaine

Roux à 37 euros. Les passionnés de la Loire voteront pour la cuvée Acacia, un pouilly-fumé signé par Didier Dagueneau à 34 euros. En rouge, discret gamay de Touraine à 19,50 euros. On vote sans hésiter pour le Château Montjouan, délicieux premières côtes de bordeaux (30 euros). Plus chic et plus cher – meilleur, aussi –, le Château Petit Figeac 98 s'offre à 60 euros. Il manque à la carte l'un ou l'autre coup de cœur pour en extraire des perles.

Parmi les autres thèmes plus spécifiques dont on traite dans cette classe de segments, on retrouve, entre autres, les produits pâtissiers (segment 3). On retrouve aussi des segments de documents traitant de la gastronomie et auxquels il est impossible d'associer des sous-thèmes plus spécifiques (segment 6)

[SEGMENT 3] La tradition n'en est pourtant pas à son premier bouleversement. Puisqu'autrefois, la bûche toute de bois venait bien du dehors pour se consumer, arrosée d'huile, de sel et de vin cuit, dans l'âtre de la cheminée. De là à se dire qu'en cédant aux sirènes alléchantes du marketing, on fait en quelque sorte revivre le passé, il n'y a qu'un petit pas que les gourmands modernes s'empressent de franchir, de plus en plus nombreux. Les consommateurs aujourd'hui ne font presque plus rien eux-mêmes, assure Marc Vandercammen, directeur du Crioc (Centre de recherche et d'information des organisations de consommateurs). Et ils s'offrent du rêve à petite dose en se tournant vers les produits exotiques. Les goûts classiques (moka, vanille, crème fraîche...) sont remplacés par des sorbets originaux. Ainsi, chez Delhaize où l'on propose depuis le 1er décembre déjà 60 000 bûches gâteaux et 200 000 glacées, la vente des produits dits pâtissiers est en progression chaque année. La hausse profite surtout aux nouveaux produits, détaille Michel Lecomte, responsable communication produits. Des mousses de thé et fruits exotiques, des bavarois myrtilles framboises ou des desserts tout chocolat. Nous offrons même cette année une bûche 75% cacao. Le pâtissier Debailleul, dont la production reste majoritairement pâtissière (65%), a purement banni de son étalage pour la première fois les gâteaux à la crème au beurre. Ils ne correspondent plus du tout au goût du jour, assure Nelly Debailleul. C'est un dessert du passé, renchérit le chef liégeois François Tonglet. Si l'on regarde l'évolution de la gastronomie ces dernières années, la bûche est restée très en retard. Cette génoise à la crème au beurre, c'est bien plus une pâtisserie de goûter qu'un dessert d'après repas, surtout de réveillon. Le cuisinier se gardera bien d'en offrir à sa carte – Je l'ai remplacée par un mille-feuilles pur chocolat accompagné d'une glace aux truffes aux arômes boisés –, mais il avoue toutefois qu'il ne résistera pas au plaisir d'en acheter une pour la déguster en famille. C'est le lendemain qu'elle est la meilleure, on la retrouve au petit déjeuner et c'est sympa. Sans doute parce que l'estomac a oublié quelque peu les excès de la veille, mais surtout parce que le biscuit a eu le temps de s'imprégner. On peut dire, si l'on veut, qu'une bûche doit « mûrir », sourit Nelly Debailleul. Nos produits sont souvent meilleurs le deuxième ou le troisième jour. A côté de trois parfums très très chocolat, une mousse au citron vert et fraise des bois voisine avec des glaces vanille-chocolat meringuée, lait d'amandes-framboises ou vanille-caramel aux noix de pécans caramélisées. Nous les vendons déjà très bien, surtout en portions individuelles, depuis le 4 décembre, ajoute-t-elle.

[SEGMENT 6] Sur la petite centaine de maisons traversées anonymement, 31 adresses ont été finalement retenues, en toute subjectivité : elles « en étaient », un point c'est tout. On y retrouve pêle-mêle quelques classiques comme l'Arcadi café, le Café Camille, Bonsoir Clara, Lola ou Le Pain et le Vin au côté de « petits nouveaux » (Un Peu beaucoup, *Eat & Love*...) qui sur la durée ont tout a prouvé. Le Belga Queen et le Café Belga témoignent d'un retour à une forme de belgitude qui n'exclut pas le plaisir que l'on peut avoir à plonger son pain dans un bol chez Mange Ta Soupe ou de s'essayer à la cuisine énergisante et quasi archéologique de Tan qui prône l'alimentation « vive ». Les cuisines exotiques ne s'invitent qu'au menu de ceux qui ont choisi de défendre la « *fusion food* » ou l'art de jongler avec la mondialisation des denrées : une tendance plaisante qui a toutefois le défaut d'uniformiser quelque peu des cartes – le livre lui-même le révèle par une lecture suivie – qui ne peuvent plus se concevoir sans sushis, tartares et carpaccios. Autant de plats qui succombent à cet autre courant, la mode du cru qu'il vous faudra apprendre à « cuisiner » chez vous. Puisque, selon nos deux observateurs qui se veulent aussi annonceurs des courants de demain, la maison comme lieu de réception va redevenir le centre du monde.

Si l'utilisateur désire poursuivre son analyse en sélectionnant le terme « Fromage », il lui est possible de découvrir une autre classe dans laquelle ce mot est présent. Il est alors dirigé vers une classe thématique (classe 24) dans laquelle l'opération d'extraction automatique des termes thématiques à identifié les termes « Maison », « Chef » et « Produit ». Cette classe est focalisée autour du concept de « Fromager » (figure 4.11) et les principaux segments qui y figurent sont les suivants :

[SEGMENT 3] Les grands fromagers parisiens se comptent sur les doigts d'une main. Androuet en fait partie. La maison a été créée en 1909. Elle présente un chiffre d'affaires de 4,5 millions d'euros par an et écoule 20 tonnes de fromages par mois. Ce chiffre colossal pourrait faire croire à une production industrielle. Or, c'est tout le contraire : Androuet est spécialisé dans la fromagerie fine. Quelque 250 références dont la moitié sont vieillies et affinés dans sa propre cave. Tous les fromages sont fabriqués au lait cru et sont fournis pour un quart par des petits producteurs régionaux. La vache s'approprie 65% des laitages, la chèvre 25 % et la brebis 10 %.

[SEGMENT 5] Aujourd'hui, Mme Delange, membre de la confrérie des Compagnons fromagers de Belgique, animera une conférence sur les fromages wallons, avec dégustation. Le 7 novembre, la pâtisserie et la chocolaterie auront leur tour. Vendredi, Benjamin Tomasetti expliquera le processus de fabrication de la bière par le biais de la plus petite brasserie artisanale du monde dont il est l'heureux concepteur.

[SEGMENT 6] La venue de deux confréries françaises, représentant les fromages Munster et Neufchâtel, invitées par les deux confréries locales (Confrérie du Herve et Seigneurie du Remoudou), traduit une volonté d'internationaliser la manifestation et de faire valoir le patchwork gustatif des produits du terroir européen.

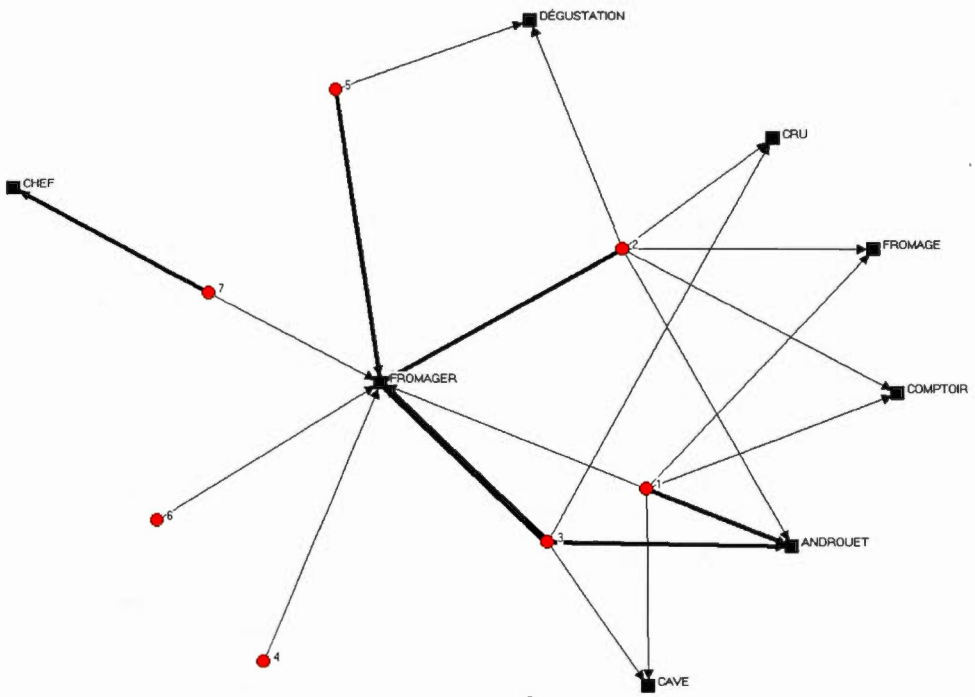


Figure 4.11. Représentation graphique du contenu thématique de la classe 24
(● = Segment et ■ = Mot).

4.6. Parcours thématique 4

Le dernier exemple d'analyse thématique que nous présentons est réalisé en employant, en premier lieu, le terme « Musulman ». L'opération d'extraction automatique des termes thématiques candidats a permis d'associer correctement ce terme thématique aux classes 18, 43, 53, 72 et 116. Chacune de ces classes traite donc directement de ce sujet, mais l'opération de classification automatique les a distinguées car, étant constituées d'un lexique légèrement différent, elles abordent cette thématique selon différentes perspectives qui leurs sont propres. Par exemple, les deux classes suivantes, toutes deux caractérisées automatiquement par le terme thématique « Musulman », ne sont pas composées du même lexique et des mêmes termes centraux (figure 4.12 et 4.13).

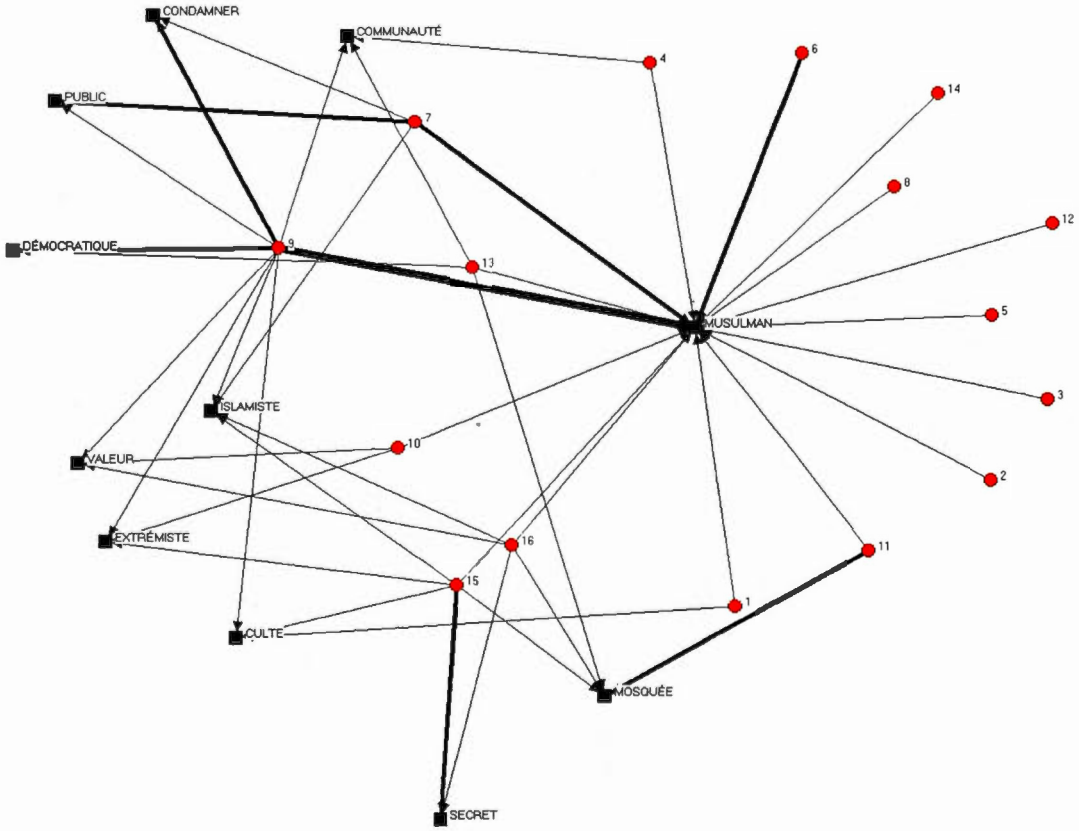


Figure 4.12. Représentation graphique du contenu thématique de la classe 43 (● = Segment et ■ = Mot).

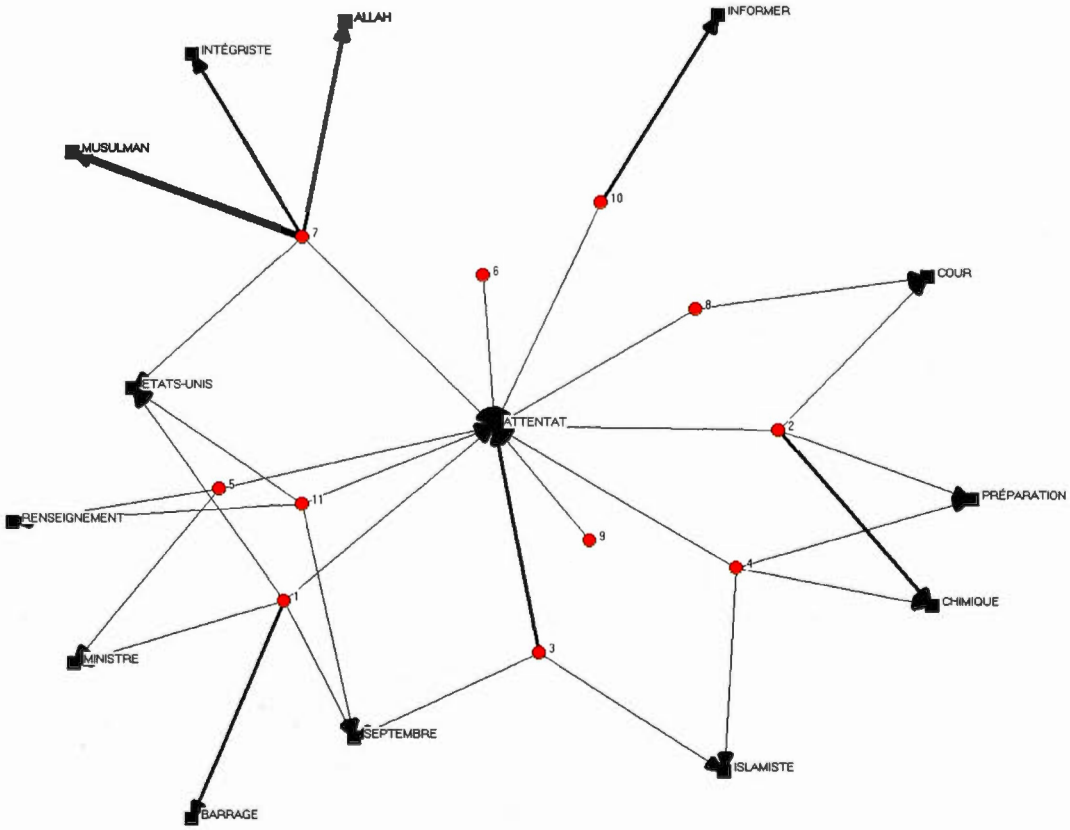


Figure 4.13. Représentation graphique du contenu thématique de la classe 18
(● = Segment et ■ = Mot).

L'utilisateur désirant consulter les documents relatant les incidents du 11 septembre 2001 privilégiera la classe 18 (dont le terme central est « Attentat ») plutôt que la classe 43 dont le terme central est « Musulman ». Il pourra cependant naviguer, en raison des termes partagés par les deux classes, d'une classe à l'autre. Si l'utilisateur s'attarde à la classe 18, il consultera entre autres les segments suivants :

[SEGMENT 2] Ils ont surtout saisi des flacons de substance suspecte ainsi qu'une combinaison NBC (nucléaire, bactériologique et chimique). Les matières (de la poudre et des éléments liquides) sont en cours d'analyse dans un laboratoire spécialisé pour déterminer leur éventuelle toxicité. Mais avant même de connaître les résultats de ces expertises, l'idée qu'un attentat chimique était peut-être en préparation a commencé de circuler.

[SEGMENT 3] La prise serait la plus importante réalisée en France depuis les attentats antiaméricains du 11 septembre 2001. Parmi les islamistes interpellés figure Mirouane Ben Ahmed. Franco-Algérien de 29 ans, il serait lié au désormais fameux groupe de Francfort qui préparait un attentat contre le marché de Noël et la cathédrale de Strasbourg en décembre 2000. Son épouse serait au nombre des personnes interpellées. Les deux derniers hommes seraient de nationalité algérienne pour l'un et marocaine pour l'autre.

[SEGMENT 4] Ces islamistes auraient séjourné en Afghanistan ainsi qu'en Tchétchénie. Ils auraient par ailleurs été en contact avec Rabah Kadri (dit Toufik), arrêté le mois dernier à Londres alors qu'un attentat chimique était semble-t-il en préparation dans le métro de la capitale britannique. Mais rien ne prouve pour l'instant leur appartenance à la mouvance Al-Qaïda d'Oussama Ben Laden.

Bien que cette classe de segments de documents aborde un thème très spécifique, elle comporte néanmoins certains termes généraux permettant à l'utilisateur de découvrir d'autres thèmes du corpus. Ainsi, grâce entre autres au mot « Service », il est possible de découvrir la classe thématique 81 dont le contenu est représenté par les termes « Juges », « Avocat » et « Client » (figure 4.14). Comme en témoignent les segments suivants, cette classe traite principalement du thème de la pédophilie en relatant différents faits de l'affaire Dutroux.

[SEGMENT 1] Il reste donc à Neufchâteau. Qui n'est plus un parquet de province comme il pouvait le penser. Il y a l'affaire des titres volés, liée au dossier Cools et qui retournera à Liège, plus tard. Il y joue déjà un duo avec le juge d'instruction Jean-Marc Connerotte. Et puis, le 9 août 1996, Lætitia Delhez se fait enlever à Bertrix. Michel Bourlet se rend sur place très rapidement. Le 13 août, Dutroux, Lelièvre et Michelle Martin sont arrêtés. Le 15 août, Dutroux lance : Je vais vous donner deux filles. Michel Bourlet libère Lætitia et Sabine Dardenne, disparues depuis 78 jours.

[SEGMENT 3] Vendredi, la dernière journée de débats a donné lieu à un nouvel incident qui témoigne de l'exaspération des parties civiles. L'avocat de Nihoul, Me Clément de Cléty, regrettait au cours de son intervention l'incident qui s'était produit le 24 octobre. L'avocat de Michel Lelièvre avait insinué que les parents des enfants étaient responsables du décès de la juge d'instruction Doutrèwe. Ces affirmations avaient provoqué une vive réaction de Jean-Denis Lejeune, menaçant l'avocat sur les marches du palais. Hier, Jean-Denis Lejeune a demandé au président Moinet de rappeler à l'ordre Me de Cléty. A ce moment-là, Me Sluzny a traité Jean-Denis Lejeune de menteur et d'imbécile. Il s'est excusé par la suite.

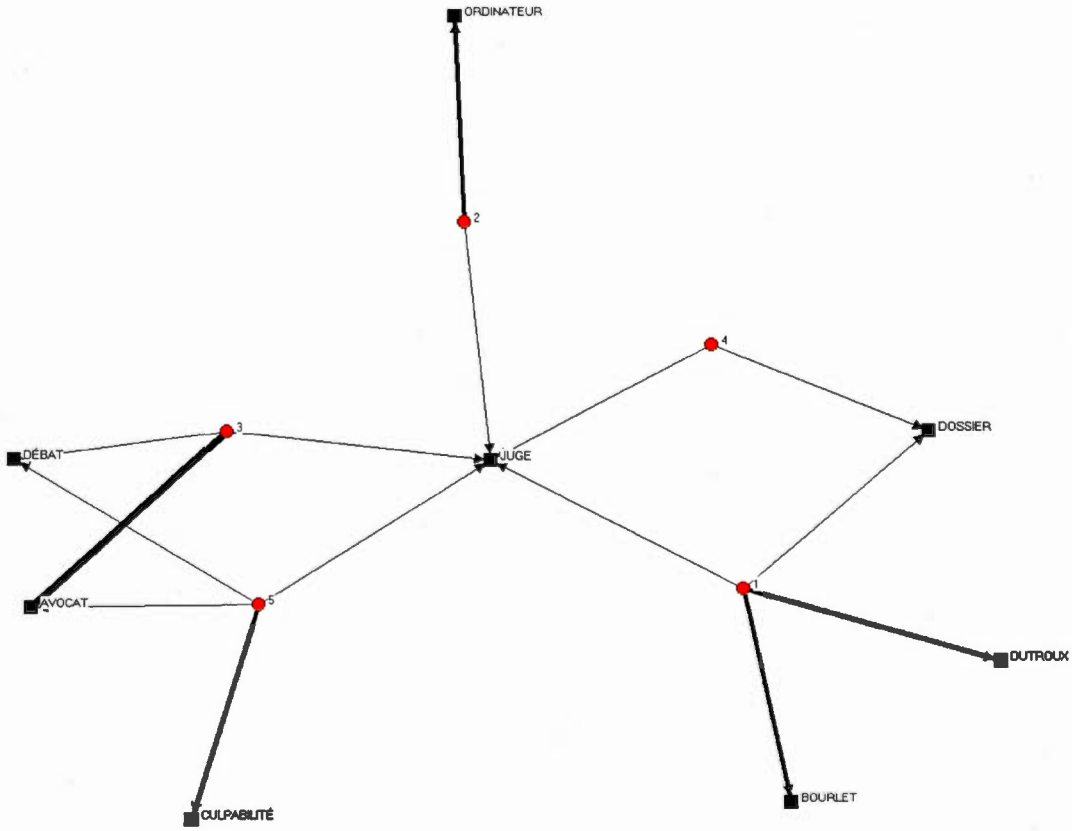


Figure 4.14. Représentation graphique du contenu thématique de la classe 81
(● = Segment et ■ = Mot).

Il n'y a évidemment pas de lien, aussi bien sémantique que pragmatique, entre ces deux thèmes reliés (le premier étant caractérisé par les termes « Attentat », « Musulman », « États-Unis », le second par les termes « Juges », « Avocat » et « Client »). D'ailleurs, dans ce projet, l'objectif de l'analyse thématique n'est pas d'identifier des liens de cette nature. Cependant, nous sommes en mesure de constater en appliquant notre démarche que le mot « Service » est présent dans ces deux regroupements thématiques différents. C'est la présence d'un même mot dans plus d'une classe qui permet le processus d'analyse thématique. Dans plusieurs cas, la présence de termes dans plus d'une classe attestera une variation sémantique de ce terme, mais il n'est pas nécessaire qu'il en soit ainsi dans tous les cas.

4.7. Remarques générales

Les exemples que nous avons présentés ne constituent qu'un très petit échantillon des nombreux parcours thématiques possibles dans notre corpus d'expérimentation. Nous croyons cependant qu'ils suffisent pour démontrer concrètement que notre démarche permet efficacement d'identifier les thèmes d'un corpus et, surtout, d'en assister informatiquement l'analyse thématique.

À la lumière de cet échantillon de résultats, deux remarques méritent d'être mentionnées. La première concerne les résultats obtenus par l'opération d'extraction automatique des termes thématiques candidats. Lors de nos expérimentations, nous avons choisi d'extraire automatiquement trois termes thématiques candidats pour chacune des classes. À cet égard, nous avons démontré que la qualité des résultats obtenus par cette opération est optimale lorsque nous ne retenons que le premier terme dont la valeur, selon la mesure d'évaluation proposée par Hirst et St-Onge, est la plus élevée. Plus nous retenons de termes, plus les performances du système tendent à se détériorer. Cependant, le fait de retenir plus d'un terme possède certains avantages non négligeables. Dans la majorité des cas, on constate que le premier terme retenu permet d'identifier le thème général de la classe dont il provient. Le fait de retenir plus d'un terme thématique engendrera une détérioration des résultats (lorsque les résultats sont comparés à l'étiquette thématique attribuée manuellement selon la mesure d'évaluation de Hirst et St-Onge) mais il permettra, en contrepartie, de préciser davantage le contenu thématique des classes. Par exemple, nous avons mentionné que les classes 18, 43, 53, 72 et 116 ont toutes été caractérisées par le terme thématique « Musulman ». Bien qu'elles partagent toutes ce terme, les classes n'abordent pas toutes ce même thème dans la même perspective. Les différents segments de ces classes n'ont pas été regroupés ensemble lors de l'opération de classification, car ils ne partagent pas un lexique majoritaire commun. En d'autres termes, les classes 18, 43, 53, 72 et 116 traitent d'un même thème général commun (caractérisé par le terme « Musulman »), mais elles l'abordent selon différentes perspectives ou elles comportent des sous-thèmes plus spécifiques dont on peut attester la présence grâce à une analyse du lexique de chaque classe. En extrayant automatiquement plus d'un terme thématique, il est donc possible de tenir compte de ces variations. Le tableau suivant représente bien la pertinence d'extraire plus d'un terme thématique pour représenter

les variations thématiques. Dans ce tableau, chaque classe est d'abord catégorisée par un même terme thématique. Ce sont les deuxième et troisième termes qui permettent d'identifier les particularités de chacune des classes.

CLASSE	MOTS THÉMATIQUES CANDIDATS
034	**Musulman**, Communauté, Ministre
043	**Musulman**, Mosquée, Islamiste
053	**Musulman**, Politique, Monde
072	**Musulman**, Islam, Communauté
116	**Musulman**, Religion, Islam

Tableau 4.1. Comparaison des termes thématiques candidats des classes 34, 43, 53, 72 et 116.

La seconde remarque découle de la première, mais elle concerne plus spécifiquement la manière dont sont catégorisés les documents de notre corpus d'expérimentation. Lorsque nous l'avons constitué, nous avons effectué des requêtes thématiques composées d'un seul mot. Ces requêtes nous ont permis de récupérer les documents dans lesquels le terme de la requête figurait comme mot-clé thématique. Cette méthode de recherche implémentée dans la version numérique du journal LE SOIR nous a permis de récupérer les documents de chacune des catégories retenues dans notre corpus d'expérimentation. Cette méthode est certes efficace, mais contrairement à ce que nous indiquent nos travaux, elle ne permet pas de distinguer les variations thématiques que l'on retrouve au sein d'une même catégorie. Comme en témoigne le tableau 4.1, les documents regroupés dans une même catégorie présentent des variations thématiques fines, mais néanmoins importantes. Ainsi, nous sommes d'avis qu'en plus d'être utile à des fins d'assistance à l'analyse thématique des documents textuels non-structurés, les méthodologies proposées dans notre projet peuvent s'avérer des plus fécondes afin d'assister le travail de catégorisation des documents à des fins de recherche et de repérage de l'information.

CONCLUSION

Dans ce travail, nous avons d'abord présenté la problématique dans laquelle s'insère notre projet de recherche (chapitre 1). À cet égard, nous avons justifié la pertinence et la nécessité des technologies visant à assister l'identification des thèmes et l'analyse thématique des documents textuels. De plus, nous avons brièvement évoqué les particularités respectives de chacune de ces tâches. Nous avons aussi présenté les principales réalisations dans ces domaines et identifié les limites des technologies existantes. Pour conclure ce chapitre, nous avons défini les objectifs spécifiques à notre projet de recherche.

Dans un second temps (chapitre 2), nous avons présenté le cadre conceptuel, ainsi que les différents travaux théoriques sur les problématiques du thème et de l'analyse thématique sur lesquels repose en partie notre démarche. À cet égard, nous vu comment notre démarche se fonde sur certaines des hypothèses théoriques développées par Kintsch et Van Dijk, Rimmon-Kenan et Rastier. Au niveau informatique, nous avons présenté la démarche méthodologique, ainsi que l'architecture informatique à laquelle nous avons eu recours dans le cadre de notre étude. Nous avons vu comment l'assistance à l'identification des thèmes et à l'analyse thématique peut être réalisée en employant certaines techniques provenant des domaines de l'intelligence artificielle et de l'apprentissage machine. Les deux principales techniques explorées dans ce projet sont la classification et la catégorisation automatiques dans leur application au traitement des documents textuels. Ces deux techniques sont complémentaires et le choix de l'une ou de l'autre est dicté principalement par la nature de l'analyse à effectuer. La catégorisation automatique est à la base d'analyses prédictives, alors que la classification automatique permet plutôt des analyses de nature exploratoire ou descriptive.

La troisième partie de ce travail (chapitre 3) a été consacrée à la présentation détaillée des résultats obtenus lors de nos expérimentations.

Dans la quatrième partie de notre travail (chapitre 4), nos efforts ont consisté à démontrer comment, sur la base des résultats présentés dans le chapitre précédent, il est possible d'assister

informatiquement les opérations d'identification des thèmes, de découverte, de parcours et d'analyse thématique d'un corpus journalistique.

La principale contribution du projet réside dans la validation de l'hypothèse selon laquelle l'identification automatique des thèmes et l'analyse thématique de documents textuels peuvent être assistées informatiquement en employant certaines techniques issues de récents travaux dans les domaines de l'intelligence artificielle et de l'apprentissage machine. À cet égard, nous avons exploré la pertinence et l'efficacité de deux techniques particulières. La première fait appel au processus de catégorisation automatique. Selon les modalités d'expérimentation spécifiées et les résultats obtenus dans notre projet, l'hypothèse selon laquelle une méthodologie fondée sur la catégorisation automatique des segments de documents peut permettre d'assister l'identification du contenu thématique et l'analyse thématique est infirmée. La seconde technique fait appel aux processus de classification automatique et d'extraction automatique des termes thématiques candidats. Les résultats obtenus par cette démarche se sont avérés des plus prometteurs. Par conséquent, ces résultats permettent de valider l'hypothèse selon laquelle une méthodologie fondée sur les opérations de classification automatique et d'extraction automatique des termes thématiques candidats peut permettre d'assister l'identification des thèmes et l'analyse thématique de documents textuels non structurés.

Nous avons aussi voulu démontrer la pertinence d'une analyse fondée sur la segmentation des documents. Cette démarche possède entre autres l'avantage d'attribuer plusieurs catégories thématiques à chaque document (ce qui est plus difficilement réalisable lorsque les documents sont traités en entier). Finalement, nous avons proposé une solution simple mais originale aux problèmes inhérents aux différentes méthodes de catégorisation employant une taxinomie de catégories thématiques.

Au niveau cognitif, nous avons exploré et mis à contribution la pertinence de certains travaux en sémantique cognitive, en linguistique textuelle et en psycholinguistique dans leur application à l'informatisation des deux processus liés à la thématique. Plus spécifiquement, nous avons démontré la pertinence des travaux de Rastier, de Rimmon-Kenan, de Kintsch et de Van Dijk à des fins d'assistance informatique à l'analyse thématique. Nous avons démontré que notre démarche informatique employant le modèle vectoriel dans son application au traitement des documents peut être théoriquement fondée sur certaines des

hypothèses et des théories soutenues par ces auteurs.

Les problèmes de l'identification des thèmes et de l'analyse thématique sont des plus complexes. Il s'agit cependant de deux tâches fondamentales dans les domaines de la Lecture et de l'Analyse de Textes Assistées par Ordinateur (LATAO) et de la Gestion Électronique des Documents (GÉD). Avec la quantité croissante de documents textuels disponibles en format numérique, la nécessité de développer des applications évoluées permettant d'assister l'analyse et la gestion du contenu informationnel des documents ne fera que s'amplifier durant les prochaines années. Nous avons donc proposé une démarche méthodologique permettant d'assister ces opérations d'identification des thèmes et d'analyse thématique. Nous sommes d'avis que les applications de notre démarche sont des plus nombreuses.

Au-delà de son application à la découverte et au parcours thématique de corpus documentaires volumineux, nous croyons que la méthodologie proposée peut être des plus fécondes pour de nombreuses tâches de gestion et d'analyse des documents textuels. Comme le soulignent Louwerse et Van Peer (2002, p. 215) :

> *One of the fundamental characteristics of themes is that they allow the grouping of meanings into manageable chunks. In this sense one could say that the function of themes is – to use a fashionable term – "knowledge management": themes render the multitude of information meaningful by streamlining individual pieces of information into meaningful whole which can then be processed more effectively [...].*

Par ailleurs, les opérations de catégorisation et de classification automatiques, en raison de leur généricité, peuvent être employées entre autres afin d'assister la recherche d'informations. D'ailleurs, certains moteurs de recherche sur Internet font déjà appel à des techniques mises en valeur dans notre projet. C'est entre autre le cas des moteurs de recherche CLUSTY et GROKKER qui appliquent certains mécanismes de classification sur les documents récupérés. L'application de cette méthodologie ne se limite pas exclusivement à la recherche et à la découverte d'informations. Certaines applications commerciales (telle l'application IDOL SERVER de la compagnie AUTONOMY) sont composées de modules de classification et de catégorisation afin d'assister la création et la mise à jour de taxinomies. Finalement, en raison des efforts continus pour développer une nouvelle génération d'Internet, le Web Sémantique, l'identification de liens thématiques entre différents documents fera certainement l'objet de travaux de recherche des plus soutenus.

Dans son état actuel, notre contribution demeure cependant modeste. Nous avons proposé une démarche méthodologique permettant d'assister deux tâches cognitivement complexes liées à l'analyse de l'information. Mais il reste encore du travail à faire avant que cette méthodologie ne puisse être déployée pleinement. Au niveau de la démarche fondée sur l'opération de catégorisation, il importe d'explorer davantage différents paramètres qui permettraient d'optimiser les résultats obtenus. Ainsi, nous sommes d'avis que l'ajout de certaines fonctionnalités visant à filtrer plus adéquatement le lexique initial du corpus permettrait d'accroître les résultats de cette opération. Aussi, l'application d'un thésaurus sur la liste des termes retenus pour générer la matrice contribuerait à un meilleur regroupement des segments de documents. Finalement, dans tout processus de catégorisation, le problème découlant de l'application d'une taxinomie demeure entier. Il importe donc que des efforts de recherche soient dirigés vers le développement d'applications visant à assister rigoureusement la construction et la mise à jour de taxinomies. Ces dernières sont fondamentales dans les domaines de l'analyse et de la gestion électronique des documents.

La démarche fondée sur la classification automatique permet de contourner certaines limites inhérentes à celles fondées sur la catégorisation. La poursuite de certains travaux permettrait néanmoins d'en augmenter davantage les performances. Nous croyons que le réseau neuronal ART1 possède de nombreux avantages pour la classification de données textuelles. Cependant, le domaine de la classification est actuellement un important lieu de recherche et il est probable que de nouvelles méthodes de classification puissent s'avérer encore plus efficaces (voir entre autres Hélie *et al.*, 2006). Par ailleurs, nous avons appliqué une méthode simple, mais néanmoins efficace, afin d'extraire automatiquement les termes thématiques candidats (TF · IDF). Plusieurs autres mesures visant à pondérer la valeur informative ou discriminante de termes ont été explorées durant les dernières années. Yang et Pedersen (1997) ont démontré la pertinence d'employer uniquement la fréquence comme mesure de pondération, mais il est possible que certaines combinaisons de mesures puissent offrir des performances supérieures.

Bref, le développement d'outils informatiques visant à assister l'identification du contenu thématique et l'analyse thématique de documents textuels demeure un lieu de recherche des plus complexes. Nous avons modestement proposé une méthodologie, fondée théoriquement, pouvant assister ces deux opérations. Cependant, les problématiques reliées à l'identification

et à l'analyse des thèmes soulèvent des enjeux théoriques et des défis techniques importants qui feront – nous l'espérons – l'objet de plusieurs initiatives de recherche durant les prochaines années.

ANNEXE 1

STATISTIQUES DE L'ENSEMBLE D'APPRENTISSAGE 1

(ne tenant pas compte de la distribution des catégories thématiques)

FEATURE	MEAN	STANDARD DEVIATION	MINIMUM	MAXIMUM	MINIMUM GIVEN	MAXIMUM GIVEN	MISSING
ARABE	0.03	0.21	0.00	3.00	0.00	4.00	0
ATTENTAT	0.01	0.11	0.00	1.00	0.00	2.00	0
CINEMA	0.03	0.22	0.00	3.00	0.00	3.00	0
CLIJSTERS	0.04	0.21	0.00	2.00	0.00	2.00	0
COLLEGE	0.01	0.14	0.00	2.00	0.00	2.00	0
COMMISSION	0.04	0.24	0.00	3.00	0.00	3.00	0
CONSEIL	0.07	0.27	0.00	2.00	0.00	3.00	0
CONSOMMATEUR	0.02	0.13	0.00	2.00	0.00	2.00	0
CREATION	0.02	0.16	0.00	2.00	0.00	2.00	0
CUISINE	0.01	0.14	0.00	3.00	0.00	3.00	0
DIRECTEUR	0.02	0.15	0.00	2.00	0.00	2.00	0
DOCTEUR	0.01	0.07	0.00	1.00	0.00	1.00	0
DROIT	0.05	0.22	0.00	2.00	0.00	2.00	0
ELECTRABEL	0.02	0.17	0.00	2.00	0.00	3.00	0
EUROS	0.10	0.56	0.00	7.00	0.00	7.00	0
FER	0.03	0.17	0.00	1.00	0.00	2.00	0
FILM	0.05	0.29	0.00	4.00	0.00	5.00	0
FOI	0.07	0.29	0.00	2.00	0.00	2.00	0
FORMATION	0.02	0.16	0.00	3.00	0.00	3.00	0
FROMAGER	0.01	0.13	0.00	3.00	0.00	3.00	0
GARE	0.10	0.42	0.00	4.00	0.00	4.00	0
GOUVERNEMENT	0.05	0.22	0.00	2.00	0.00	3.00	0
GUERRE	0.02	0.18	0.00	4.00	0.00	4.00	0
HENIN	0.03	0.18	0.00	2.00	0.00	2.00	0
INFORMATIQUE	0.02	0.16	0.00	2.00	0.00	5.00	0
INTERNET	0.01	0.12	0.00	2.00	0.00	3.00	0
IRAK	0.01	0.10	0.00	2.00	0.00	2.00	0
ISLAMIQUE	0.02	0.13	0.00	1.00	0.00	2.00	0
ISLAMISTE	0.03	0.24	0.00	5.00	0.00	5.00	0

Feature	Mean	Standard deviation	Minimum	Maximum	Minimum given	Maximum given	Missing
JUGE	0.03	0.18	0.00	3.00	0.00	3.00	0
JUSTICE	0.03	0.17	0.00	2.00	0.00	2.00	0
LABORATOIRE	0.01	0.14	0.00	2.00	0.00	2.00	0
LIGUE	0.01	0.09	0.00	1.00	0.00	3.00	0
LOGICIEL	0.03	0.18	0.00	2.00	0.00	3.00	0
LOI	0.03	0.20	0.00	2.00	0.00	3.00	0
MALADE	0.01	0.14	0.00	2.00	0.00	3.00	0
MASTERS	0.01	0.12	0.00	1.00	0.00	1.00	0
MATCH	0.03	0.18	0.00	2.00	0.00	2.00	0
MOSQUEE	0.01	0.14	0.00	3.00	0.00	3.00	0
MUSULMAN	0.04	0.28	0.00	4.00	0.00	4.00	0
MUSULMANE	0.01	0.12	0.00	1.00	0.00	3.00	0
MEDECIN	0.06	0.34	0.00	6.00	0.00	6.00	0
MEDECINE	0.02	0.22	0.00	5.00	0.00	5.00	0
MEDICAL	0.03	0.19	0.00	3.00	0.00	4.00	0
NUCLEAIRE	0.02	0.15	0.00	2.00	0.00	2.00	0
NUMERIQUE	0.01	0.09	0.00	1.00	0.00	2.00	0
ORDINATEUR	0.02	0.18	0.00	2.00	0.00	2.00	0
PEINE	0.02	0.15	0.00	1.00	0.00	1.00	0
PLAINTE	0.02	0.17	0.00	2.00	0.00	2.00	0
POLICE	0.04	0.22	0.00	3.00	0.00	4.00	0
POLITIQUE	0.06	0.28	0.00	3.00	0.00	3.00	0
POMME	0.01	0.12	0.00	2.00	0.00	2.00	0
PORTO	0.03	0.19	0.00	3.00	0.00	3.00	0
PRISON	0.02	0.19	0.00	3.00	0.00	3.00	0
PROCUREUR	0.03	0.18	0.00	2.00	0.00	2.00	0
PRODUCTEUR	0.02	0.13	0.00	2.00	0.00	3.00	0
PROFESSEUR	0.02	0.16	0.00	2.00	0.00	2.00	0
PROFESSIONNEL	0.03	0.17	0.00	2.00	0.00	2.00	0
RECETTE	0.01	0.13	0.00	3.00	0.00	3.00	0
RELIGIEUX	0.02	0.13	0.00	1.00	0.00	1.00	0
RELIGION	0.02	0.16	0.00	2.00	0.00	3.00	0
REALISATEUR	0.01	0.13	0.00	3.00	0.00	3.00	0
RESEAU	0.05	0.27	0.00	3.00	0.00	3.00	0
SANTE	0.04	0.22	0.00	3.00	0.00	3.00	0
SCIENTIFIQUE	0.02	0.14	0.00	2.00	0.00	2.00	0
SET	0.02	0.17	0.00	3.00	0.00	3.00	0
SNCB	0.04	0.22	0.00	3.00	0.00	3.00	0
SPORT	0.01	0.12	0.00	1.00	0.00	1.00	0
SPORTIF	0.02	0.15	0.00	3.00	0.00	3.00	0
SYSTEME	0.04	0.23	0.00	3.00	0.00	3.00	0

Feature	Mean	Standard deviation	Minimum	Maximum	Minimum given	Maximum given	Missing
TECHNOLOGIE	0.01	0.10	0.00	1.00	0.00	3.00	0
TENNIS	0.02	0.17	0.00	3.00	0.00	3.00	0
TERRAIN	0.02	0.13	0.00	2.00	0.00	2.00	0
TOURNOI	0.04	0.21	0.00	2.00	0.00	2.00	0
TRAIN	0.04	0.27	0.00	4.00	0.00	4.00	0
TRIBUNAL	0.02	0.14	0.00	2.00	0.00	2.00	0
UNIVERSITAIRE	0.02	0.13	0.00	2.00	0.00	2.00	0
UNIVERSITE	0.02	0.19	0.00	4.00	0.00	4.00	0
VAINQUEUR	0.02	0.14	0.00	2.00	0.00	2.00	0
VICTIME	0.02	0.14	0.00	1.00	0.00	2.00	0
VICTOIRE	0.03	0.21	0.00	3.00	0.00	3.00	0
VIN	0.02	0.18	0.00	4.00	0.00	4.00	0
VIOLENCE	0.02	0.18	0.00	2.00	0.00	2.00	0
VELO	0.02	0.14	0.00	2.00	0.00	5.00	0
ELECTRIQUE	0.03	0.17	0.00	2.00	0.00	2.00	0
ELEVE	0.01	0.12	0.00	2.00	0.00	2.00	0
ETUDE	0.03	0.19	0.00	2.00	0.00	2.00	0
ETUDIER	0.07	0.32	0.00	4.00	0.00	4.00	0

Number of input features : 88
Number of classes : 10
Number of cases : 1083
Cases with missing values : 0

Classes

CINEMA : 72 cases
ELECTRICITE : 97 cases
GARE : 108 cases
GASTRONOMIE : 47 cases
INFORMATIQUE : 90 cases
ISLAM : 165 cases
MEDECIN : 136 cases
PEDOPHILIE : 106 cases
TENNIS : 157 cases
UNIVERSITE : 105 cases

ANNEXE 2

STATISTIQUES DE L'ENSEMBLE DE TEST 1

(ne tenant pas compte de la distribution des catégories thématiques)

FEATURE	MEAN	STANDARD DEVIATION	MINIMUM	MAXIMUM	MINIMUM GIVEN	MAXIMUM GIVEN	MISSING
ARABE	0.03	0.23	0.23	4.00	0.00	4.00	0
ATTENTAT	0.02	0.15	0.15	2.00	0.00	2.00	0
CINEMA	0.04	0.22	0.22	2.00	0.00	3.00	0
CLIJSTERS	0.04	0.21	0.21	2.00	0.00	2.00	0
COLLEGE	0.02	0.15	0.15	2.00	0.00	2.00	0
COMMISSION	0.02	0.16	0.16	2.00	0.00	3.00	0
CONSEIL	0.07	0.28	0.28	3.00	0.00	3.00	0
CONSOMMATEUR	0.02	0.15	0.15	2.00	0.00	2.00	0
CREATION	0.02	0.15	0.15	2.00	0.00	2.00	0
CUISINE	0.01	0.10	0.10	1.00	0.00	3.00	0
DIRECTEUR	0.03	0.17	0.17	1.00	0.00	2.00	0
DOCTEUR	0.02	0.13	0.13	1.00	0.00	1.00	0
DROIT	0.07	0.28	0.28	2.00	0.00	2.00	0
ELECTRABEL	0.02	0.21	0.21	3.00	0.00	3.00	0
EUROS	0.07	0.44	0.44	6.00	0.00	7.00	0
FER	0.04	0.21	0.21	2.00	0.00	2.00	0
FILM	0.08	0.37	0.37	5.00	0.00	5.00	0
FOI	0.07	0.28	0.28	2.00	0.00	2.00	0
FORMATION	0.03	0.20	0.20	2.00	0.00	3.00	0
FROMAGER	0.02	0.20	0.20	3.00	0.00	3.00	0
GARE	0.08	0.35	0.35	4.00	0.00	4.00	0
GOUVERNEMENT	0.06	0.26	0.26	3.00	0.00	3.00	0
GUERRE	0.03	0.24	0.24	4.00	0.00	4.00	0
HENIN	0.03	0.17	0.17	2.00	0.00	2.00	0
INFORMATIQUE	0.04	0.30	0.30	5.00	0.00	5.00	0
INTERNET	0.02	0.20	0.20	3.00	0.00	3.00	0
IRAK	0.01	0.12	0.12	2.00	0.00	2.00	0
ISLAMIQUE	0.02	0.16	0.16	2.00	0.00	2.00	0
ISLAMISTE	0.02	0.16	0.16	2.00	0.00	5.00	0

Feature	Mean	Standard deviation	Minimum	Maximum	Minimum given	Maximum given	Missing
JUGE	0.01	0.10	0.10	1.00	0.00	3.00	0
JUSTICE	0.02	0.13	0.13	1.00	0.00	2.00	0
LABORATOIRE	0.02	0.14	0.14	2.00	0.00	2.00	0
LIGUE	0.02	0.18	0.18	3.00	0.00	3.00	0
LOGICIEL	0.02	0.19	0.19	3.00	0.00	3.00	0
LOI	0.03	0.21	0.21	3.00	0.00	3.00	0
MALADE	0.01	0.16	0.16	3.00	0.00	3.00	0
MASTERS	0.02	0.14	0.14	1.00	0.00	1.00	0
MATCH	0.02	0.15	0.15	1.00	0.00	2.00	0
MOSQUEE	0.01	0.14	0.14	2.00	0.00	3.00	0
MUSULMAN	0.06	0.37	0.37	4.00	0.00	4.00	0
MUSULMANE	0.03	0.20	0.20	3.00	0.00	3.00	0
MEDECIN	0.07	0.28	0.28	2.00	0.00	6.00	0
MEDECINE	0.01	0.11	0.11	2.00	0.00	5.00	0
MEDICAL	0.04	0.26	0.26	4.00	0.00	4.00	0
NUCLEAIRE	0.01	0.11	0.11	2.00	0.00	2.00	0
NUMERIQUE	0.01	0.13	0.13	2.00	0.00	2.00	0
ORDINATEUR	0.02	0.15	0.15	2.00	0.00	2.00	0
PEINE	0.02	0.15	0.15	1.00	0.00	1.00	0
PLAINTE	0.02	0.15	0.15	2.00	0.00	2.00	0
POLICE	0.02	0.23	0.23	4.00	0.00	4.00	0
POLITIQUE	0.09	0.35	0.35	3.00	0.00	3.00	0
POMME	0.01	0.11	0.11	2.00	0.00	2.00	0
PORTO	0.02	0.14	0.14	1.00	0.00	3.00	0
PRISON	0.02	0.15	0.15	2.00	0.00	3.00	0
PROCUREUR	0.02	0.15	0.15	1.00	0.00	2.00	0
PRODUCTEUR	0.04	0.21	0.21	3.00	0.00	3.00	0
PROFESSEUR	0.04	0.20	0.20	2.00	0.00	2.00	0
PROFESSIONNEL	0.02	0.13	0.13	1.00	0.00	2.00	0
RECETTE	0.01	0.10	0.10	1.00	0.00	3.00	0
RELIGIEUX	0.03	0.17	0.17	1.00	0.00	1.00	0
RELIGION	0.02	0.19	0.19	3.00	0.00	3.00	0
REALISATEUR	0.01	0.11	0.11	1.00	0.00	3.00	0
RESEAU	0.06	0.29	0.29	3.00	0.00	3.00	0
SANTE	0.04	0.26	0.26	3.00	0.00	3.00	0
SCIENTIFIQUE	0.01	0.10	0.10	1.00	0.00	2.00	0
SET	0.02	0.18	0.18	3.00	0.00	3.00	0
SNCB	0.03	0.19	0.19	2.00	0.00	3.00	0
SPORT	0.01	0.10	0.10	1.00	0.00	1.00	0
SPORTIF	0.01	0.10	0.10	1.00	0.00	3.00	0
SYSTEME	0.04	0.22	0.22	2.00	0.00	3.00	0

FEATURE	MEAN	STANDARD DEVIATION	MINIMUM	MAXIMUM	MINIMUM GIVEN	MAXIMUM GIVEN	MISSING
TECHNOLOGIE	0.02	0.18	0.18	3.00	0.00	3.00	0
TENNIS	0.04	0.18	0.18	1.00	0.00	3.00	0
TERRAIN	0.02	0.13	0.13	1.00	0.00	2.00	0
TOURNOI	0.04	0.22	0.22	2.00	0.00	2.00	0
TRAIN	0.02	0.20	0.20	3.00	0.00	4.00	0
TRIBUNAL	0.01	0.11	0.11	1.00	0.00	2.00	0
UNIVERSITAIRE	0.01	0.10	0.10	1.00	0.00	2.00	0
UNIVERSITE	0.01	0.10	0.10	1.00	0.00	4.00	0
VAINQUEUR	0.02	0.13	0.13	1.00	0.00	2.00	0
VICTIME	0.03	0.17	0.17	2.00	0.00	2.00	0
VICTOIRE	0.04	0.20	0.20	2.00	0.00	3.00	0
VIN	0.01	0.13	0.13	2.00	0.00	4.00	0
VIOLENCE	0.02	0.14	0.14	2.00	0.00	2.00	0
VELO	0.02	0.26	0.26	5.00	0.00	5.00	0
ELECTRIQUE	0.03	0.19	0.19	2.00	0.00	2.00	0
ELEVE	0.02	0.16	0.16	2.00	0.00	2.00	0
ETUDE	0.01	0.12	0.12	1.00	0.00	2.00	0
ETUDIER	0.06	0.28	0.28	2.00	0.00	4.00	0

Number of input features : 88
Number of classes : 10
Number of cases : 542
Cases with missing values : 0

Classes

CINEMA : 37 cases
ELECTRICITE : 43 cases
GARE : 45 cases
GASTRONOMIE : 29 cases
INFORMATIQUE : 54 cases
ISLAM : 91 cases
MEDECIN : 71 cases
PEDOPHILIE : 44 cases
TENNIS : 77 cases
UNIVERSITE : 51 cases

ANNEXE 3

STATISTIQUES DE L'ENSEMBLE D'APPRENTISSAGE 2

(tenant compte de la distribution des catégories thématiques)

Feature	Mean	Standard deviation	Minimum	Maximum	Minimum given	Maximum given	Missing
ARABE	0.03	0.21	0.00	3.00	0.00	4.00	0
ATTENTAT	0.01	0.11	0.00	1.00	0.00	2.00	0
CINEMA	0.03	0.22	0.00	3.00	0.00	3.00	0
CLIJSTERS	0.04	0.21	0.00	2.00	0.00	2.00	0
COLLEGE	0.01	0.13	0.00	2.00	0.00	2.00	0
COMMISSION	0.04	0.23	0.00	3.00	0.00	3.00	0
CONSEIL	0.06	0.26	0.00	2.00	0.00	3.00	0
CONSOMMATEUR	0.02	0.15	0.00	2.00	0.00	2.00	0
CREATION	0.02	0.16	0.00	2.00	0.00	2.00	0
CUISINE	0.01	0.15	0.00	3.00	0.00	3.00	0
DIRECTEUR	0.02	0.15	0.00	2.00	0.00	2.00	0
DOCTEUR	0.01	0.08	0.00	1.00	0.00	1.00	0
DROIT	0.05	0.23	0.00	2.00	0.00	2.00	0
ELECTRABEL	0.02	0.17	0.00	2.00	0.00	3.00	0
EUROS	0.10	0.56	0.00	7.00	0.00	7.00	0
FER	0.03	0.17	0.00	1.00	0.00	2.00	0
FILM	0.05	0.29	0.00	4.00	0.00	5.00	0
FOI	0.07	0.30	0.00	2.00	0.00	2.00	0
FORMATION	0.02	0.17	0.00	3.00	0.00	3.00	0
FROMAGER	0.01	0.15	0.00	3.00	0.00	3.00	0
GARE	0.10	0.41	0.00	4.00	0.00	4.00	0
GOUVERNEMENT	0.05	0.24	0.00	3.00	0.00	3.00	0
GUERRE	0.02	0.18	0.00	4.00	0.00	4.00	0
HENIN	0.03	0.18	0.00	2.00	0.00	2.00	0
INFORMATIQUE	0.03	0.18	0.00	2.00	0.00	5.00	0
INTERNET	0.01	0.12	0.00	2.00	0.00	3.00	0
IRAK	0.01	0.11	0.00	2.00	0.00	2.00	0
ISLAMIQUE	0.02	0.13	0.00	1.00	0.00	2.00	0
ISLAMISTE	0.03	0.24	0.00	5.00	0.00	5.00	0

Feature	Mean	Standard deviation	Minimum	Maximum	Minimum given	Maximum given	Missing
JUGE	0.02	0.18	0.00	3.00	0.00	3.00	0
JUSTICE	0.03	0.18	0.00	2.00	0.00	2.00	0
LABORATOIRE	0.01	0.14	0.00	2.00	0.00	2.00	0
LIGUE	0.01	0.09	0.00	1.00	0.00	3.00	0
LOGICIEL	0.03	0.18	0.00	2.00	0.00	3.00	0
LOI	0.03	0.20	0.00	2.00	0.00	3.00	0
MALADE	0.01	0.14	0.00	2.00	0.00	3.00	0
MASTERS	0.01	0.12	0.00	1.00	0.00	1.00	0
MATCH	0.03	0.18	0.00	2.00	0.00	2.00	0
MOSQUEE	0.01	0.14	0.00	3.00	0.00	3.00	0
MUSULMAN	0.04	0.28	0.00	4.00	0.00	4.00	0
MUSULMANE	0.01	0.12	0.00	1.00	0.00	3.00	0
MEDECIN	0.06	0.34	0.00	6.00	0.00	6.00	0
MEDECINE	0.02	0.22	0.00	5.00	0.00	5.00	0
MEDICAL	0.03	0.19	0.00	3.00	0.00	4.00	0
NUCLEAIRE	0.02	0.15	0.00	2.00	0.00	2.00	0
NUMERIQUE	0.01	0.09	0.00	1.00	0.00	2.00	0
ORDINATEUR	0.03	0.19	0.00	2.00	0.00	2.00	0
PEINE	0.02	0.14	0.00	1.00	0.00	1.00	0
PLAINTE	0.02	0.17	0.00	2.00	0.00	2.00	0
POLICE	0.03	0.21	0.00	3.00	0.00	4.00	0
POLITIQUE	0.06	0.28	0.00	3.00	0.00	3.00	0
POMME	0.01	0.12	0.00	2.00	0.00	2.00	0
PORTO	0.03	0.19	0.00	3.00	0.00	3.00	0
PRISON	0.02	0.19	0.00	3.00	0.00	3.00	0
PROCUREUR	0.03	0.18	0.00	2.00	0.00	2.00	0
PRODUCTEUR	0.02	0.13	0.00	2.00	0.00	3.00	0
PROFESSEUR	0.02	0.16	0.00	2.00	0.00	2.00	0
PROFESSIONNEL	0.03	0.17	0.00	2.00	0.00	2.00	0
RECETTE	0.01	0.13	0.00	3.00	0.00	3.00	0
RELIGIEUX	0.02	0.13	0.00	1.00	0.00	1.00	0
RELIGION	0.02	0.16	0.00	2.00	0.00	3.00	0
REALISATEUR	0.01	0.13	0.00	3.00	0.00	3.00	0
RESEAU	0.05	0.27	0.00	3.00	0.00	3.00	0
SANTE	0.03	0.22	0.00	3.00	0.00	3.00	0
SCIENTIFIQUE	0.02	0.14	0.00	2.00	0.00	2.00	0
SET	0.02	0.17	0.00	3.00	0.00	3.00	0
SNCB	0.04	0.22	0.00	3.00	0.00	3.00	0
SPORT	0.01	0.12	0.00	1.00	0.00	1.00	0
SPORTIF	0.02	0.16	0.00	3.00	0.00	3.00	0
SYSTEME	0.05	0.24	0.00	3.00	0.00	3.00	0

Feature	Mean	Standard deviation	Minimum	Maximum	Minimum given	Maximum given	Missing
TECHNOLOGIE	0.01	0.10	0.00	1.00	0.00	3.00	0
TENNIS	0.02	0.17	0.00	3.00	0.00	3.00	0
TERRAIN	0.02	0.13	0.00	2.00	0.00	2.00	0
TOURNOI	0.04	0.21	0.00	2.00	0.00	2.00	0
TRAIN	0.04	0.27	0.00	4.00	0.00	4.00	0
TRIBUNAL	0.01	0.12	0.00	2.00	0.00	2.00	0
UNIVERSITAIRE	0.02	0.13	0.00	2.00	0.00	2.00	0
UNIVERSITE	0.02	0.19	0.00	4.00	0.00	4.00	0
VAINQUEUR	0.02	0.14	0.00	2.00	0.00	2.00	0
VICTIME	0.02	0.14	0.00	1.00	0.00	2.00	0
VICTOIRE	0.03	0.21	0.00	3.00	0.00	3.00	0
VIN	0.02	0.19	0.00	4.00	0.00	4.00	0
VIOLENCE	0.02	0.18	0.00	2.00	0.00	2.00	0
VELO	0.02	0.14	0.00	2.00	0.00	5.00	0
ELECTRIQUE	0.03	0.17	0.00	2.00	0.00	2.00	0
ELEVE	0.01	0.12	0.00	2.00	0.00	2.00	0
ETUDE	0.03	0.19	0.00	2.00	0.00	2.00	0
ETUDIER	0.07	0.31	0.00	4.00	0.00	4.00	0

Number of input features : 88
Number of classes : 10
Number of cases : 1084
Cases with missing values : 0

Classes

CINEMA : 73 cases
ELECTRICITE : 93 cases
GARE : 102 cases
GASTRONOMIE : 51 cases
INFORMATIQUE : 96 cases
ISLAM : 171 cases
MEDECIN : 138 cases
PEDOPHILIE : 100 cases
TENNIS : 156 cases
UNIVERSITE : 104 cases

ANNEXE 4

STATISTIQUES DE L'ENSEMBLE DE TEST 2

(tenant compte de la distribution des catégories thématiques)

Feature	Mean	Standard deviation	Minimum	Maximum	Minimum given	Maximum given	Missing
ARABE	0.03	0.23	0.00	4.00	0.00	4.00	0
ATTENTAT	0.02	0.15	0.00	2.00	0.00	2.00	0
CINEMA	0.04	0.22	0.00	2.00	0.00	3.00	0
CLIJSTERS	0.04	0.21	0.00	2.00	0.00	2.00	0
COLLEGE	0.02	0.16	0.00	2.00	0.00	2.00	0
COMMISSION	0.03	0.18	0.00	2.00	0.00	3.00	0
CONSEIL	0.07	0.29	0.00	3.00	0.00	3.00	0
CONSOMMATEUR	0.01	0.10	0.00	1.00	0.00	2.00	0
CREATION	0.02	0.15	0.00	2.00	0.00	2.00	0
CUISINE	0.01	0.10	0.00	1.00	0.00	3.00	0
DIRECTEUR	0.03	0.17	0.00	1.00	0.00	2.00	0
DOCTEUR	0.01	0.12	0.00	1.00	0.00	1.00	0
DROIT	0.06	0.27	0.00	2.00	0.00	2.00	0
ELECTRABEL	0.02	0.21	0.00	3.00	0.00	3.00	0
EUROS	0.07	0.44	0.00	6.00	0.00	7.00	0
FER	0.04	0.21	0.00	2.00	0.00	2.00	0
FILM	0.08	0.37	0.00	5.00	0.00	5.00	0
FOI	0.07	0.27	0.00	2.00	0.00	2.00	0
FORMATION	0.03	0.20	0.00	2.00	0.00	3.00	0
FROMAGER	0.01	0.17	0.00	3.00	0.00	3.00	0
GARE	0.08	0.36	0.00	4.00	0.00	4.00	0
GOUVERNEMENT	0.05	0.22	0.00	2.00	0.00	3.00	0
GUERRE	0.03	0.23	0.00	4.00	0.00	4.00	0
HENIN	0.03	0.18	0.00	2.00	0.00	2.00	0
INFORMATIQUE	0.03	0.28	0.00	5.00	0.00	5.00	0
INTERNET	0.02	0.20	0.00	3.00	0.00	3.00	0
IRAK	0.01	0.11	0.00	2.00	0.00	2.00	0
ISLAMIQUE	0.02	0.16	0.00	2.00	0.00	2.00	0
ISLAMISTE	0.02	0.16	0.00	2.00	0.00	5.00	0

Feature	Mean	Standard deviation	Minimum	Maximum	Minimum given	Maximum given	Missing
JUGE	0.01	0.11	0.00	1.00	0.00	3.00	0
JUSTICE	0.01	0.12	0.00	1.00	0.00	2.00	0
LABORATOIRE	0.02	0.14	0.00	2.00	0.00	2.00	0
LIGUE	0.02	0.18	0.00	3.00	0.00	3.00	0
LOGICIEL	0.02	0.18	0.00	3.00	0.00	3.00	0
LOI	0.03	0.21	0.00	3.00	0.00	3.00	0
MALADE	0.01	0.16	0.00	3.00	0.00	3.00	0
MASTERS	0.02	0.14	0.00	1.00	0.00	1.00	0
MATCH	0.02	0.15	0.00	1.00	0.00	2.00	0
MOSQUEE	0.01	0.14	0.00	2.00	0.00	3.00	0
MUSULMAN	0.06	0.38	0.00	4.00	0.00	4.00	0
MUSULMANE	0.03	0.20	0.00	3.00	0.00	3.00	0
MEDECIN	0.06	0.27	0.00	2.00	0.00	6.00	0
MEDECINE	0.01	0.11	0.00	2.00	0.00	5.00	0
MEDICAL	0.04	0.27	0.00	4.00	0.00	4.00	0
NUCLEAIRE	0.01	0.11	0.00	2.00	0.00	2.00	0
NUMERIQUE	0.01	0.13	0.00	2.00	0.00	2.00	0
ORDINATEUR	0.01	0.11	0.00	1.00	0.00	2.00	0
PEINE	0.03	0.16	0.00	1.00	0.00	1.00	0
PLAINTE	0.02	0.15	0.00	2.00	0.00	2.00	0
POLICE	0.03	0.25	0.00	4.00	0.00	4.00	0
POLITIQUE	0.09	0.35	0.00	3.00	0.00	3.00	0
POMME	0.01	0.11	0.00	2.00	0.00	2.00	0
PORTO	0.02	0.14	0.00	1.00	0.00	3.00	0
PRISON	0.02	0.14	0.00	2.00	0.00	3.00	0
PROCUREUR	0.02	0.15	0.00	1.00	0.00	2.00	0
PRODUCTEUR	0.04	0.21	0.00	3.00	0.00	3.00	0
PROFESSEUR	0.04	0.20	0.00	2.00	0.00	2.00	0
PROFESSIONNEL	0.02	0.13	0.00	1.00	0.00	2.00	0
RECETTE	0.01	0.10	0.00	1.00	0.00	3.00	0
RELIGIEUX	0.03	0.17	0.00	1.00	0.00	1.00	0
RELIGION	0.02	0.19	0.00	3.00	0.00	3.00	0
REALISATEUR	0.01	0.11	0.00	1.00	0.00	3.00	0
RESEAU	0.06	0.29	0.00	3.00	0.00	3.00	0
SANTE	0.04	0.27	0.00	3.00	0.00	3.00	0
SCIENTIFIQUE	0.01	0.10	0.00	1.00	0.00	2.00	0
SET	0.02	0.18	0.00	3.00	0.00	3.00	0
SNCB	0.03	0.20	0.00	2.00	0.00	3.00	0
SPORT	0.01	0.10	0.00	1.00	0.00	1.00	0
SPORTIF	0.01	0.09	0.00	1.00	0.00	3.00	0
SYSTEME	0.03	0.19	0.00	2.00	0.00	3.00	0

Feature	Mean	Standard deviation	Minimum	Maximum	Minimum given	Maximum given	Missing
TECHNOLOGIE	0.02	0.18	0.00	3.00	0.00	3.00	0
TENNIS	0.04	0.18	0.00	1.00	0.00	3.00	0
TERRAIN	0.02	0.13	0.00	1.00	0.00	2.00	0
TOURNOI	0.04	0.22	0.00	2.00	0.00	2.00	0
TRAIN	0.03	0.20	0.00	3.00	0.00	4.00	0
TRIBUNAL	0.02	0.14	0.00	2.00	0.00	2.00	0
UNIVERSITAIRE	0.01	0.10	0.00	1.00	0.00	2.00	0
UNIVERSITE	0.01	0.10	0.00	1.00	0.00	4.00	0
VAINQUEUR	0.02	0.13	0.00	1.00	0.00	2.00	0
VICTIME	0.03	0.18	0.00	2.00	0.00	2.00	0
VICTOIRE	0.04	0.20	0.00	2.00	0.00	3.00	0
VIN	0.01	0.11	0.00	2.00	0.00	4.00	0
VIOLENCE	0.02	0.14	0.00	2.00	0.00	2.00	0
VELO	0.02	0.26	0.00	5.00	0.00	5.00	0
ELECTRIQUE	0.03	0.20	0.00	2.00	0.00	2.00	0
ELEVE	0.02	0.16	0.00	2.00	0.00	2.00	0
ETUDE	0.01	0.12	0.00	1.00	0.00	2.00	0
ETUDIER	0.06	0.28	0.00	2.00	0.00	4.00	0

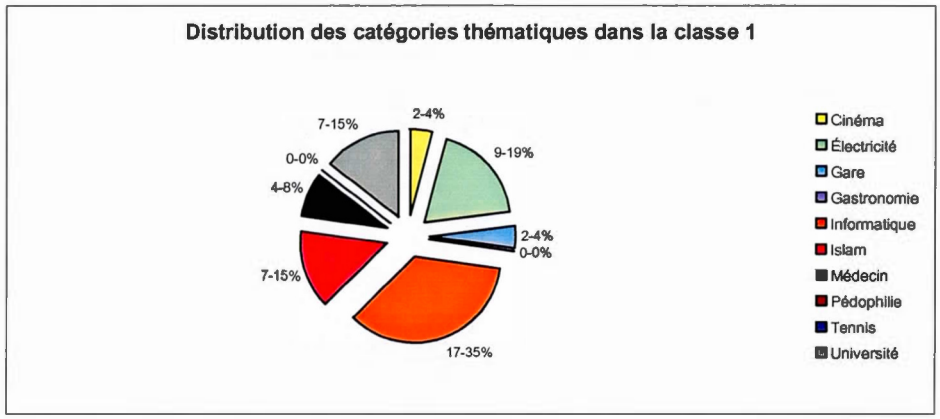
Number of input features : 88
Number of classes : 10
Number of cases : 541
Cases with missing values : 0

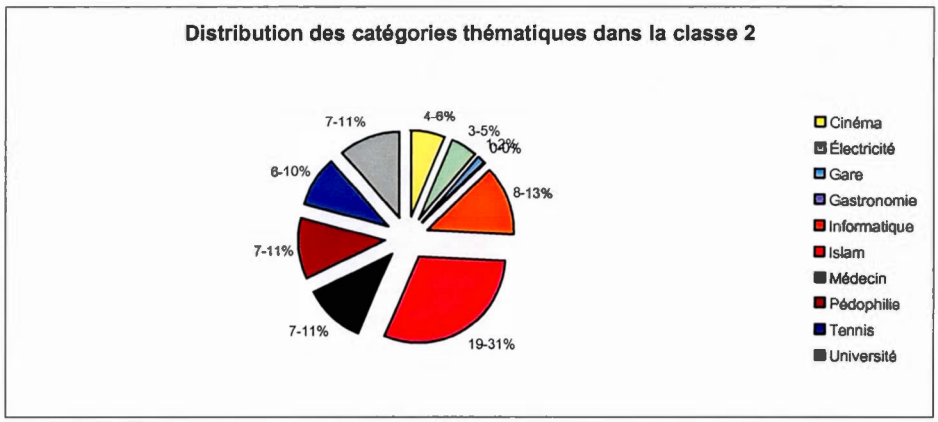
Classes

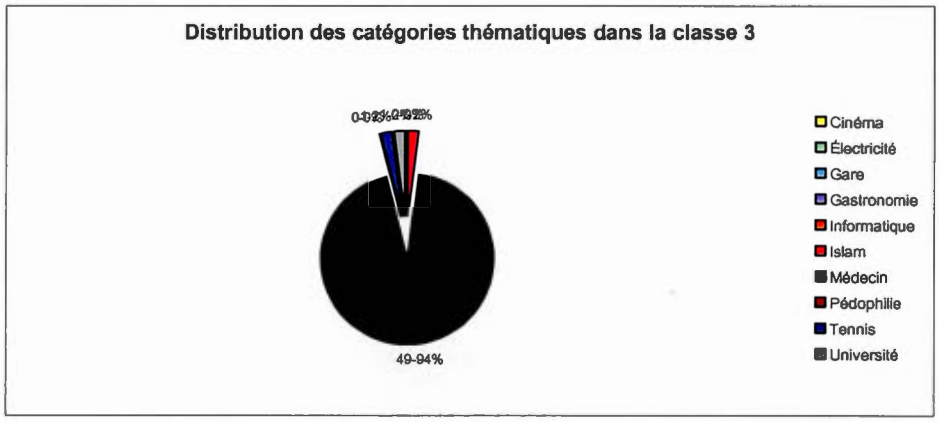
CINEMA : 36 cases
ELECTRICITE : 47 cases
GARE : 51 cases
GASTRONOMIE : 25 cases
INFORMATIQUE : 48 cases
ISLAM : 85 cases
MEDECIN : 69 cases
PEDOPHILIE : 50 cases
TENNIS : 78 cases
UNIVERSITE : 52 cases

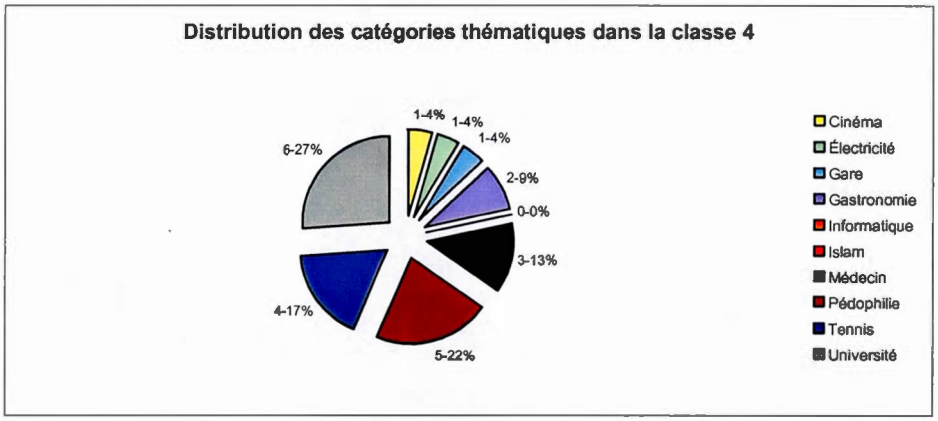
ANNEXE 5

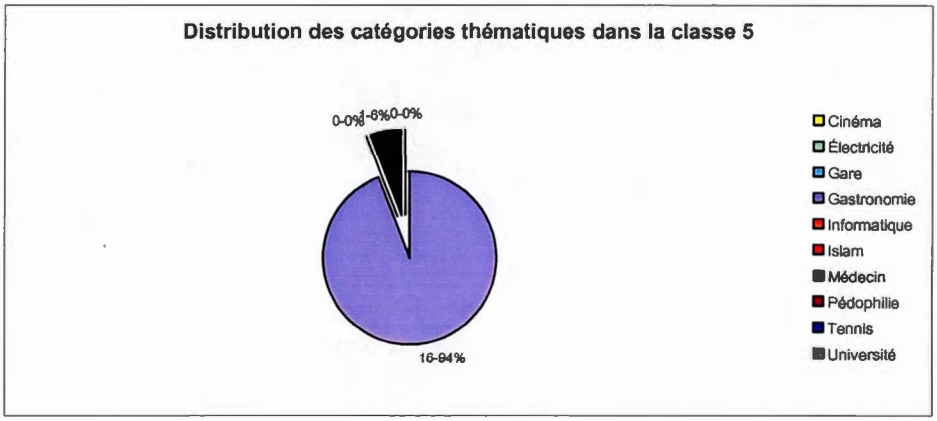
DISTRIBUTION DES CATÉGORIES THÉMATIQUES DANS CHAQUE CLASSE

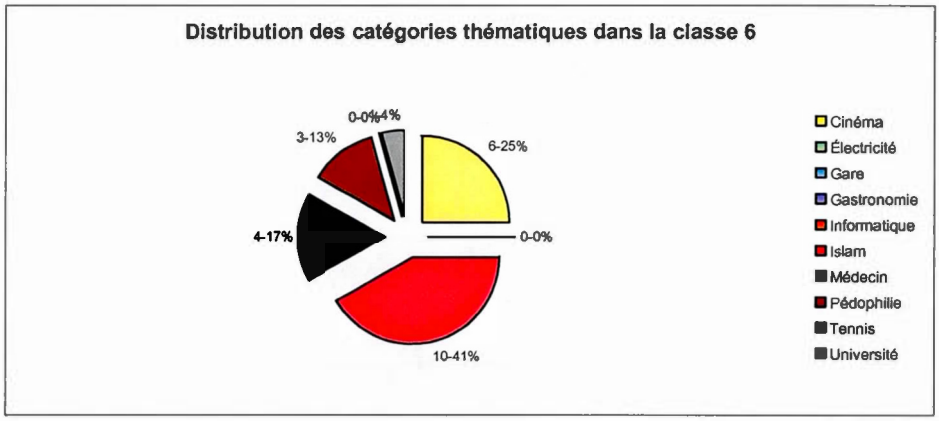


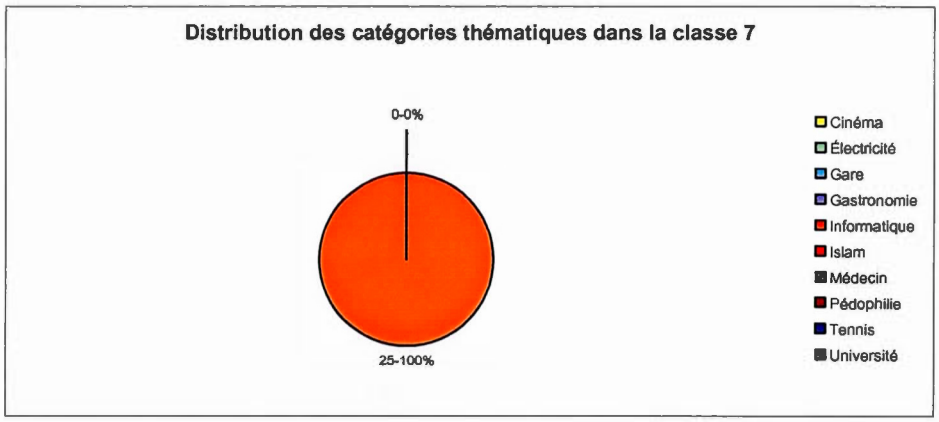


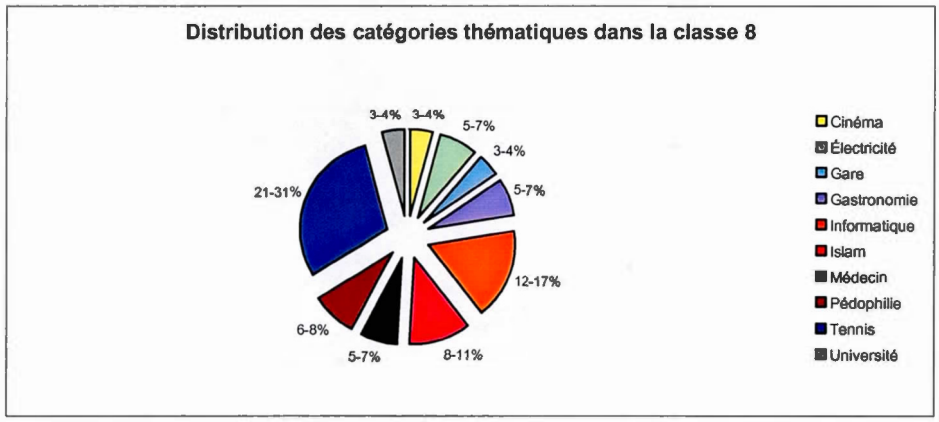

Distribution des catégories thématiques dans la classe 3
49-94%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

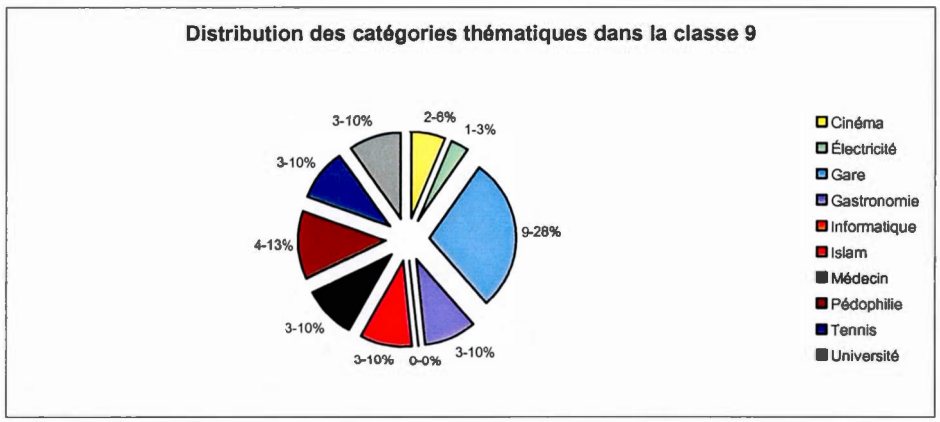

Distribution des catégories thématiques dans la classe 4
1-4%
1-4%
1-4%
2-9%
0-0%
3-13%
5-22%
4-17%
6-27%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

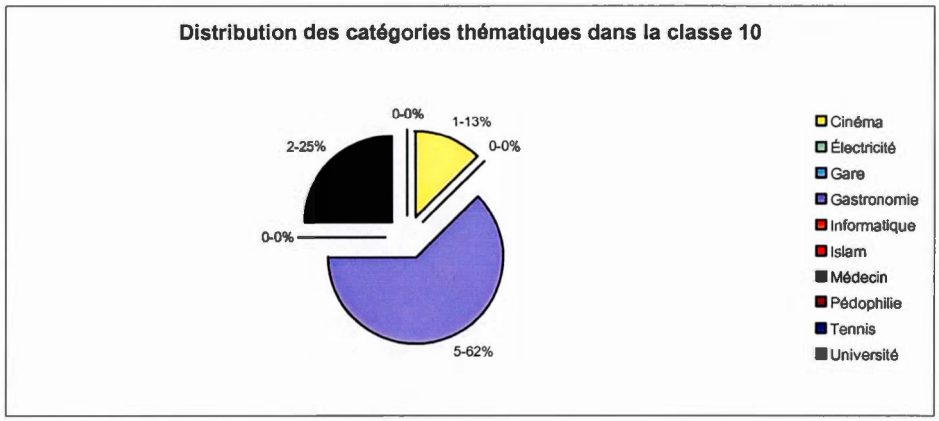

Distribution des catégories thématiques dans la classe 5
0-0%
1-6%
0-0%
16-94%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

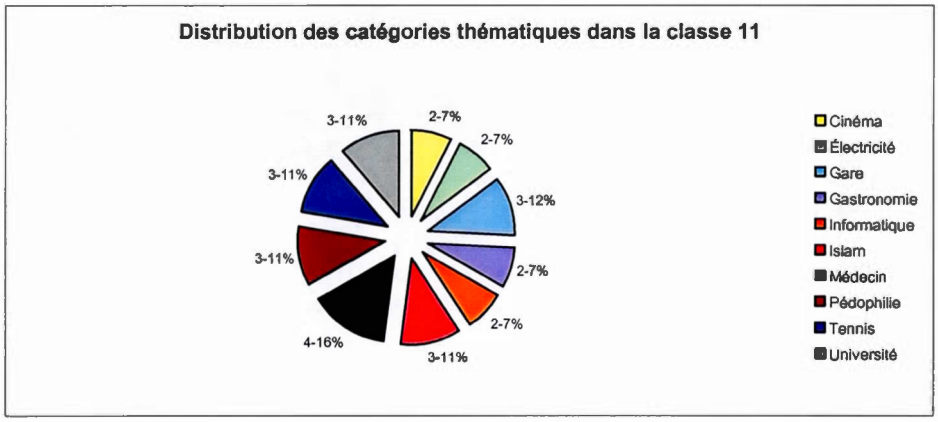

Distribution des catégories thématiques dans la classe 6
0-0%
4%
3-13%
6-25%
4-17%
0-0%
10-41%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

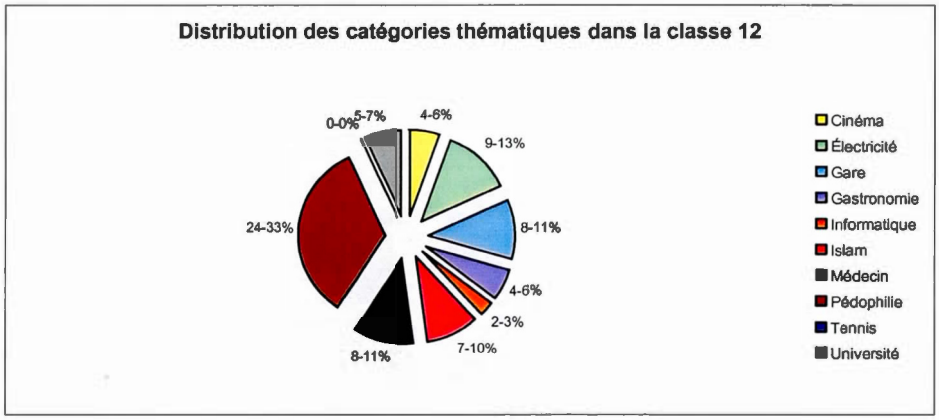

Distribution des catégories thématiques dans la classe 7
0-0%
25-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

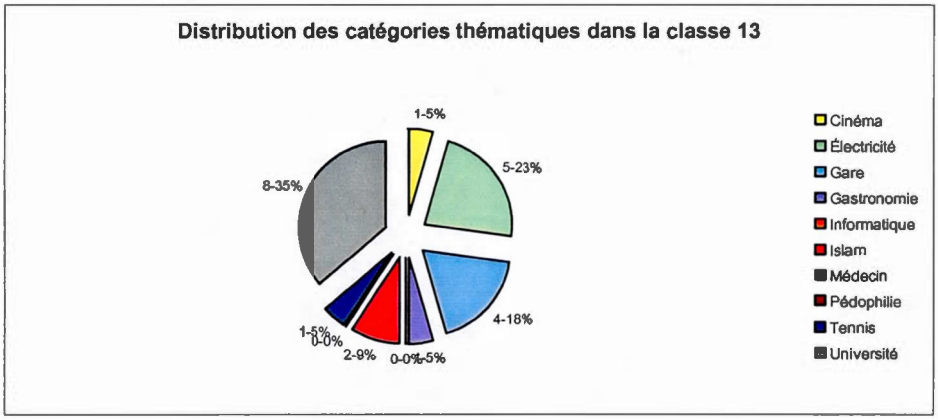

Distribution des catégories thématiques dans la classe 8
3-4%
3-4%
5-7%
3-4%
5-7%
21-31%
12-17%
6-8%
5-7%
8-11%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

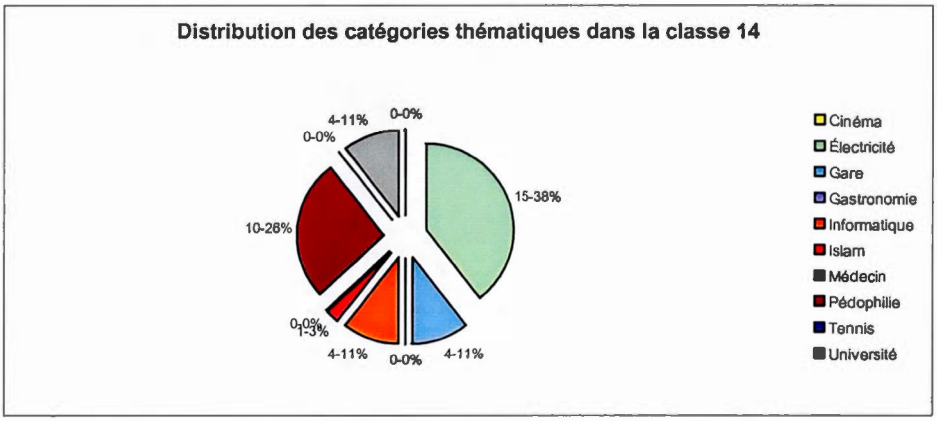

Distribution des catégories thématiques dans la classe 9
3-10%
2-6%
1-3%
3-10%
9-28%
4-13%
3-10%
3-10%
0-0%
3-10%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

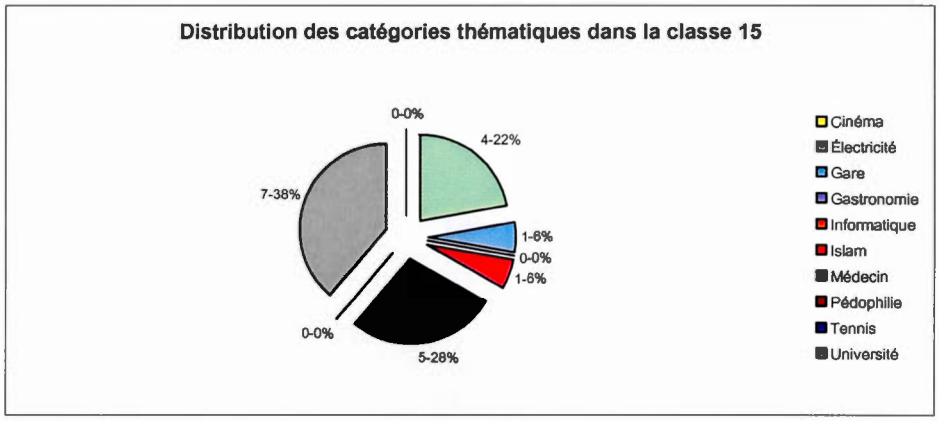

Distribution des catégories thématiques dans la classe 10
0-0%
1-13%
2-25%
0-0%
0-0%
5-62%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

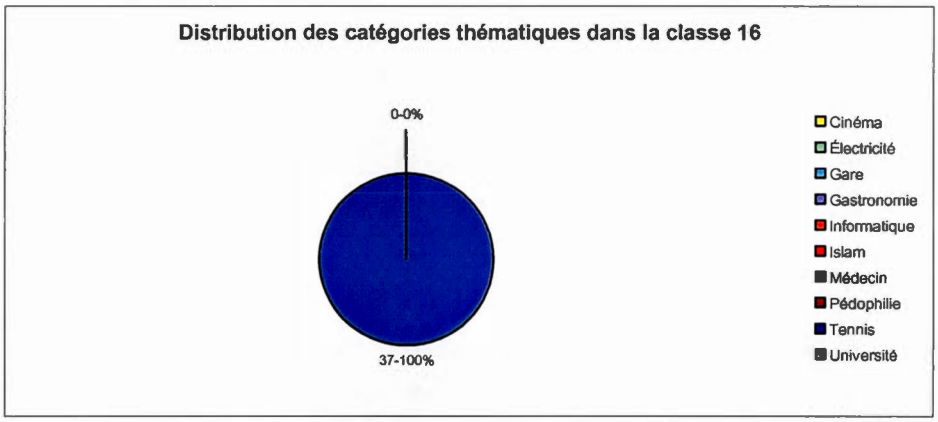

Distribution des catégories thématiques dans la classe 11
3-11%
2-7%
2-7%
3-11%
3-12%
3-11%
2-7%
2-7%
4-16%
3-11%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

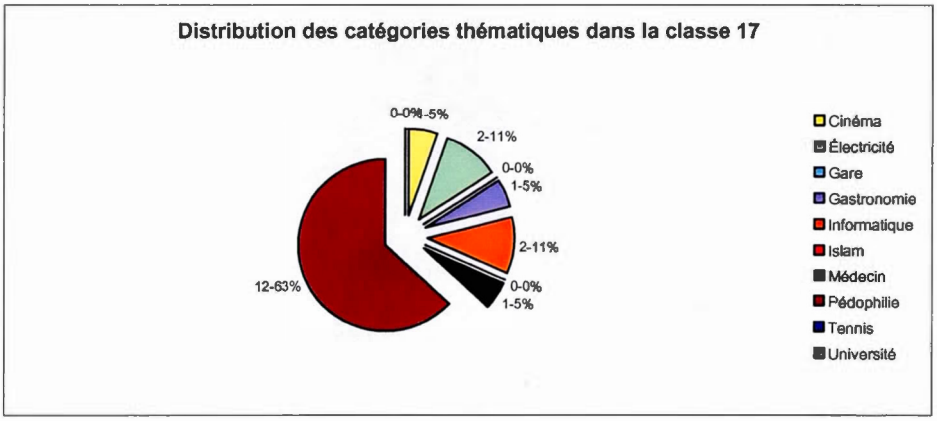

Distribution des catégories thématiques dans la classe 12
0-0%
5-7%
4-6%
9-13%
8-11%
4-6%
2-3%
7-10%
8-11%
24-33%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 13
1-5%
5-23%
4-18%
0-0%
1-5%
2-9%
0-0%
1-5%
8-35%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

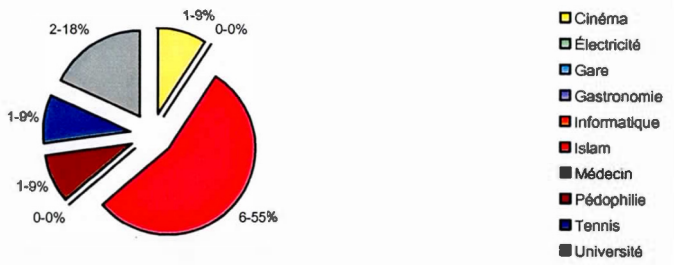

Distribution des catégories thématiques dans la classe 14
4-11%
0-0%
0-0%
15-38%
10-26%
0-0%
1-3%
4-11%
0-0%
4-11%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 15
0-0%
4-22%
7-38%
1-6%
0-0%
1-6%
0-0%
5-28%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 16
0-0%
37-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

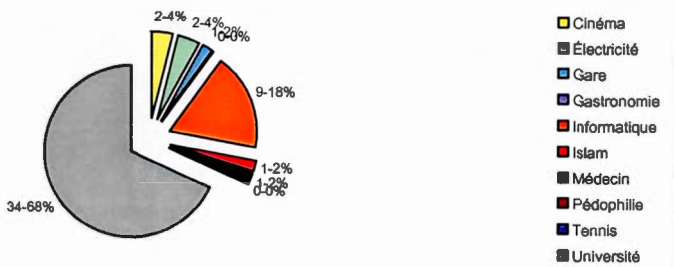

Distribution des catégories thématiques dans la classe 17
0-0%
5%
2-11%
0-0%
1-5%
2-11%
12-63%
0-0%
1-5%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

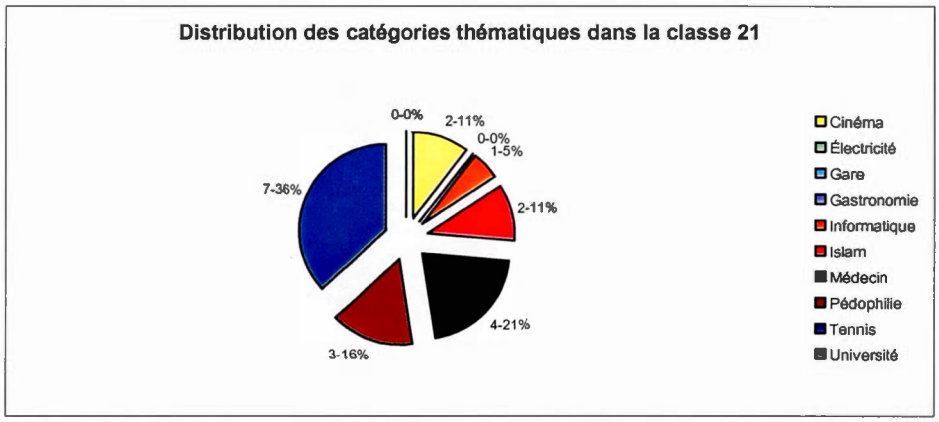
Distribution des catégories thématiques dans la classe 18

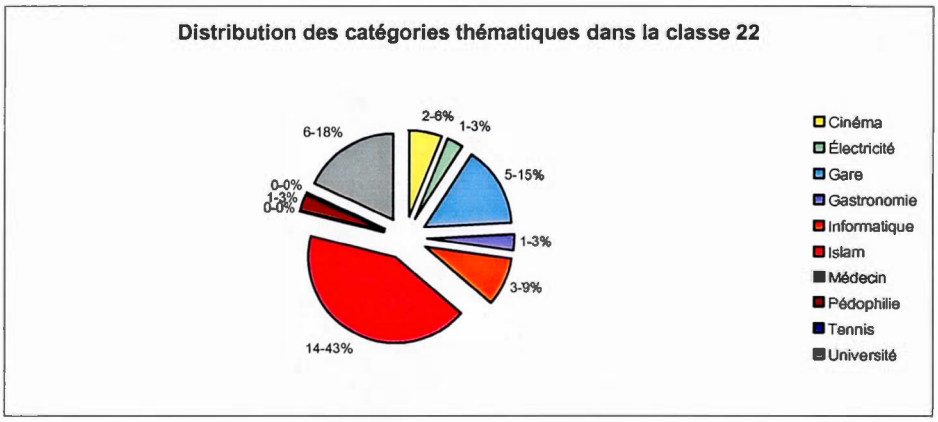


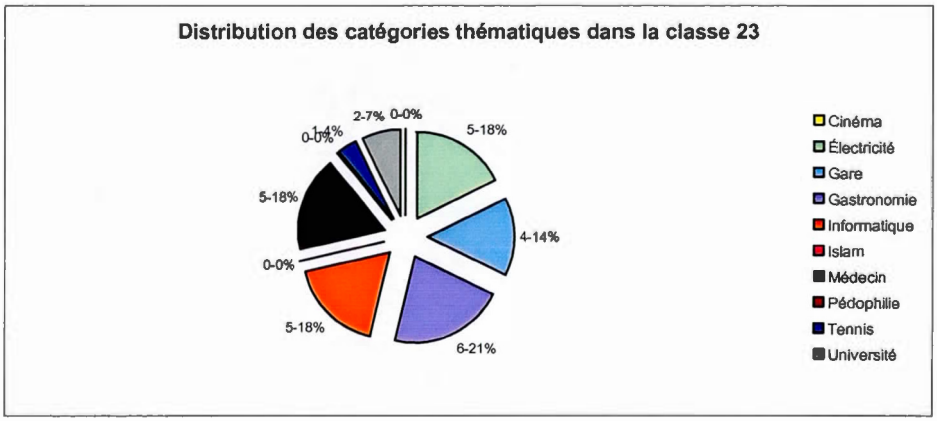
Distribution des catégories thématiques dans la classe 19

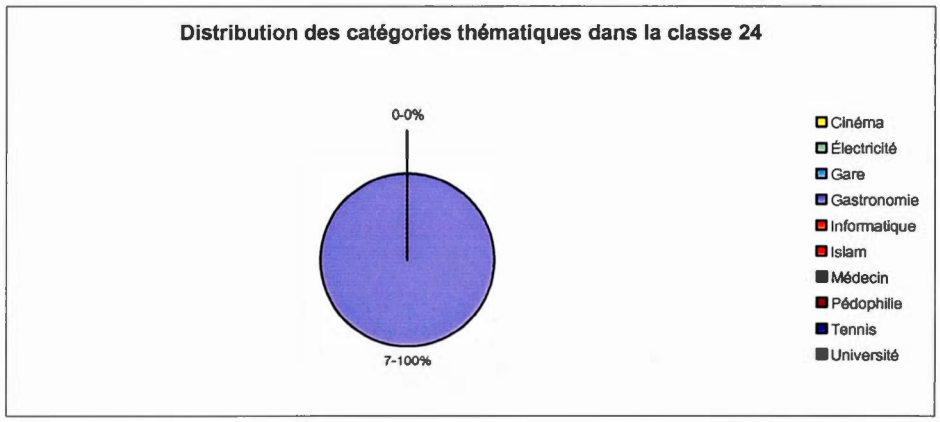
Distribution des catégories thématiques dans la classe 20

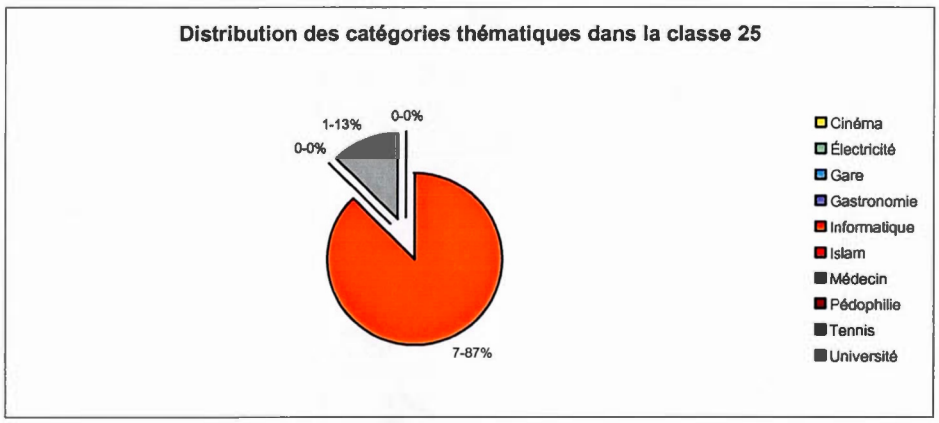


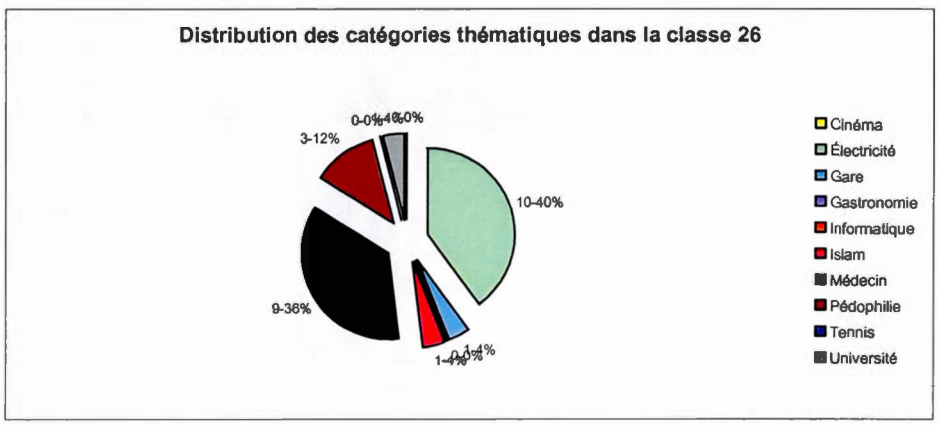

Distribution des catégories thématiques dans la classe 21
0-0%
2-11%
0-0%
1-5%
2-11%
4-21%
3-16%
7-36%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

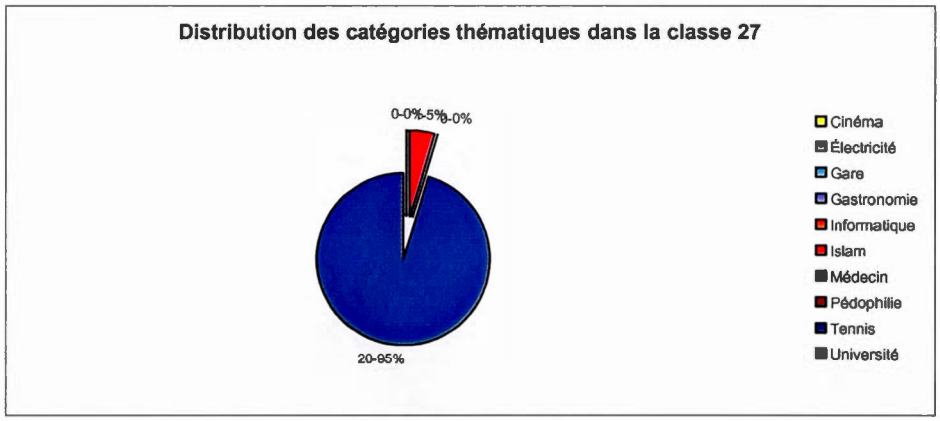

Distribution des catégories thématiques dans la classe 22
2-6%
1-3%
5-15%
1-3%
3-9%
14-43%
0-0%
1-3%
0-0%
6-18%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

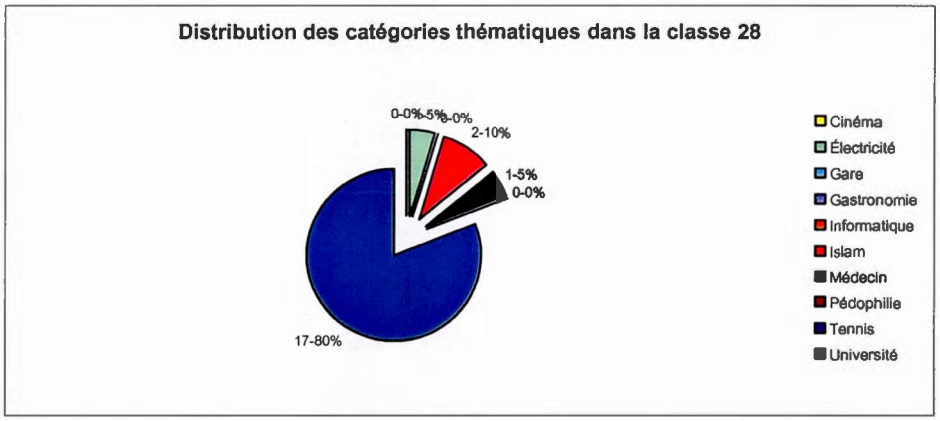

Distribution des catégories thématiques dans la classe 23
2-7%
0-0%
1-4%
0-0%
5-18%
4-14%
6-21%
5-18%
0-0%
5-18%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

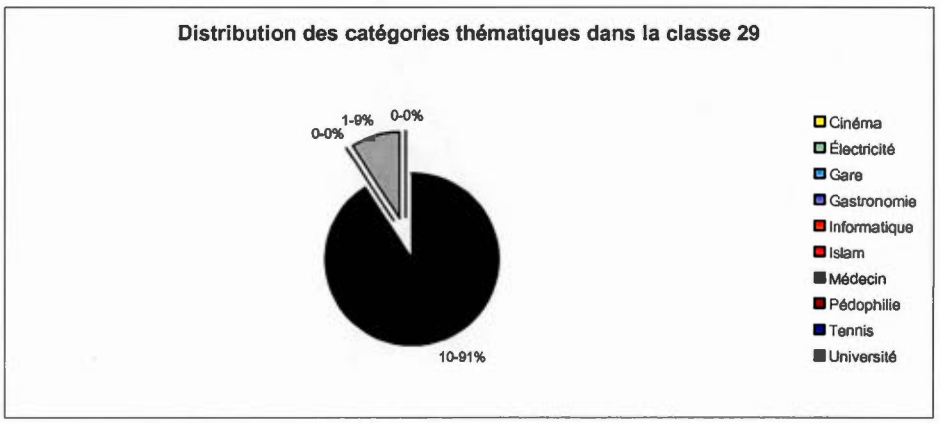

Distribution des catégories thématiques dans la classe 24
0-0%
7-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

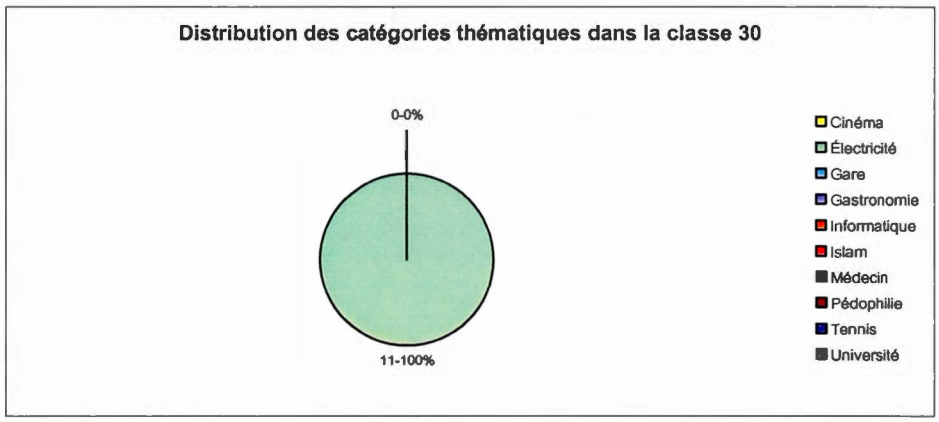

Distribution des catégories thématiques dans la classe 25
1-13%
0-0%
0-0%
7-87%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

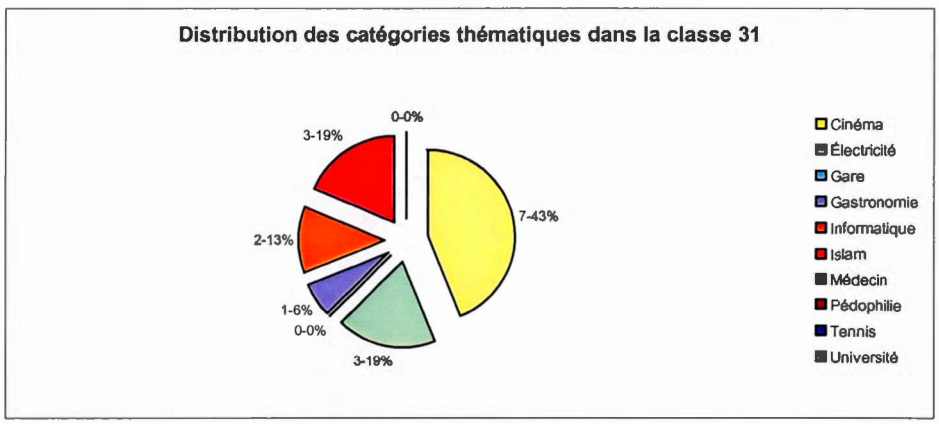

Distribution des catégories thématiques dans la classe 26
0-0%
4-0%
3-12%
10-40%
9-36%
1-4%
0-0%
1-4%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

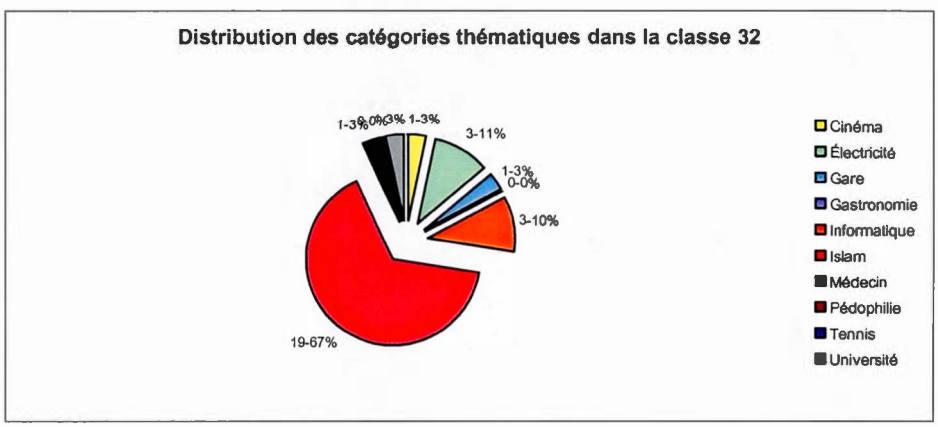

Distribution des catégories thématiques dans la classe 27
0-0%
5-5%
0-0%
20-95%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

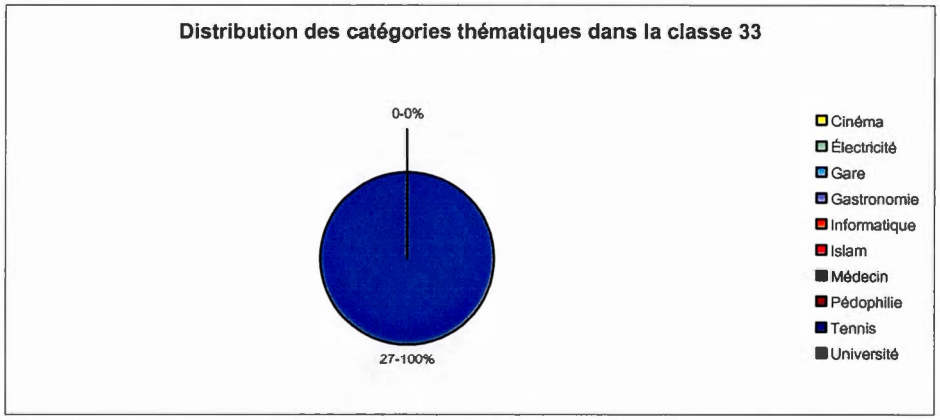

Distribution des catégories thématiques dans la classe 28
0-0%
5-5%
0-0%
2-10%
1-5%
0-0%
17-80%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

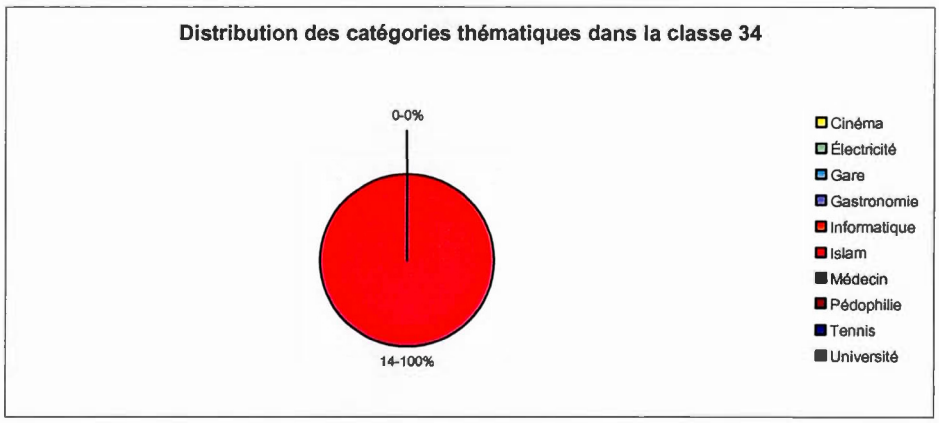

Distribution des catégories thématiques dans la classe 29
1-9%
0-0%
0-0%
10-91%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

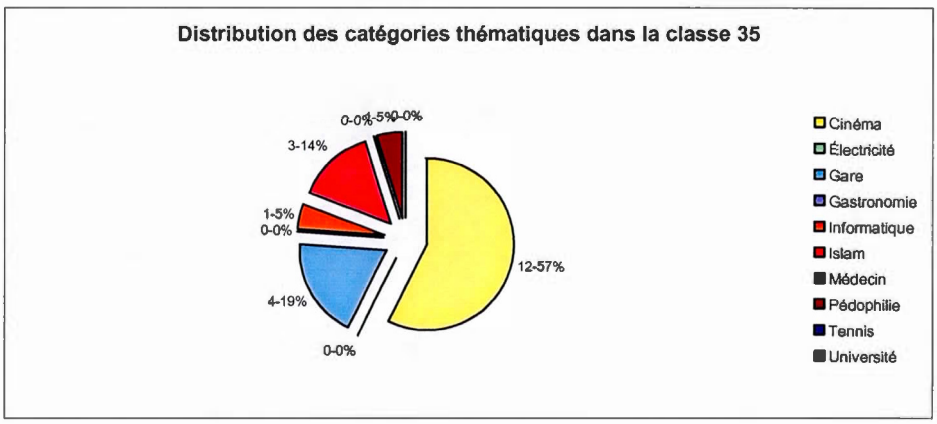

Distribution des catégories thématiques dans la classe 30
0-0%
11-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

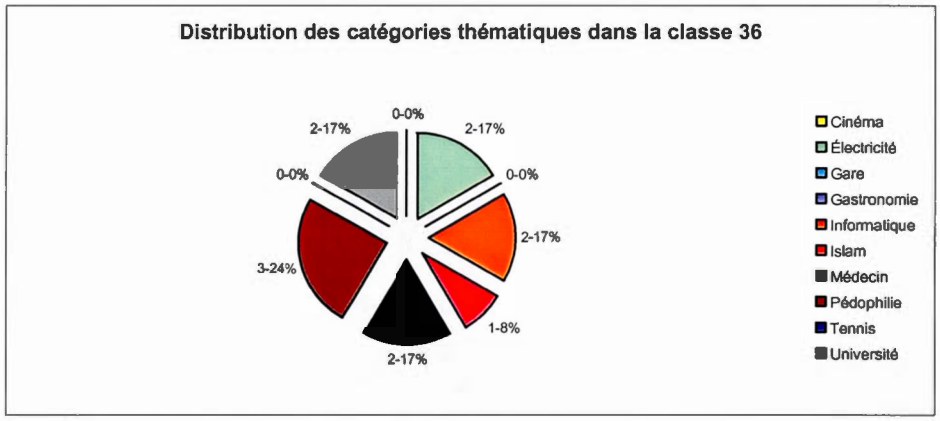

Distribution des catégories thématiques dans la classe 31
0-0%
3-19%
7-43%
2-13%
1-6%
0-0%
3-19%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

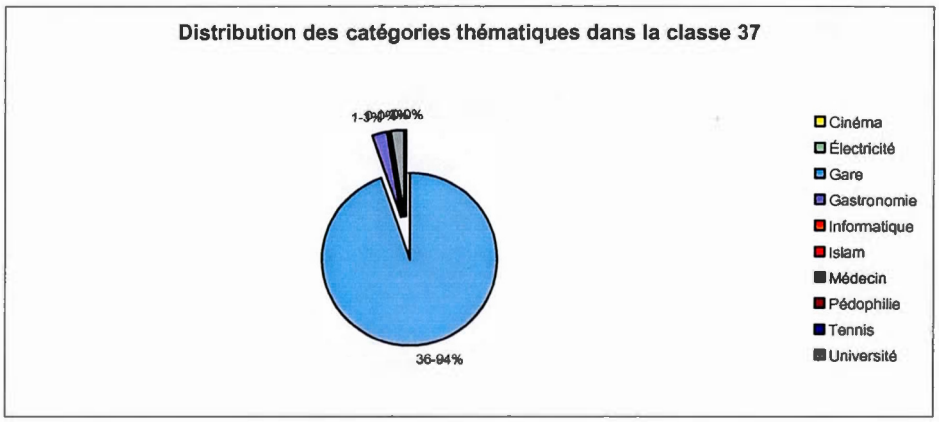

Distribution des catégories thématiques dans la classe 32
1-3%
0-0%
3-11%
1-3%
0-0%
3-10%
19-67%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

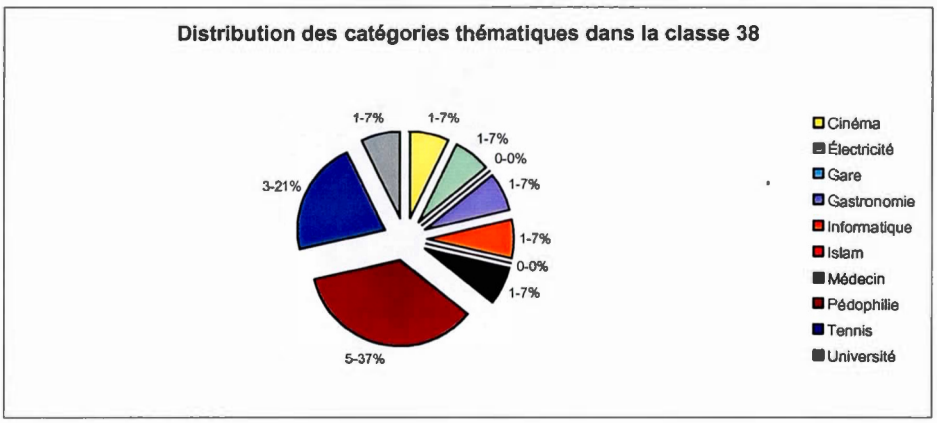

Distribution des catégories thématiques dans la classe 33
0-0%
27-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

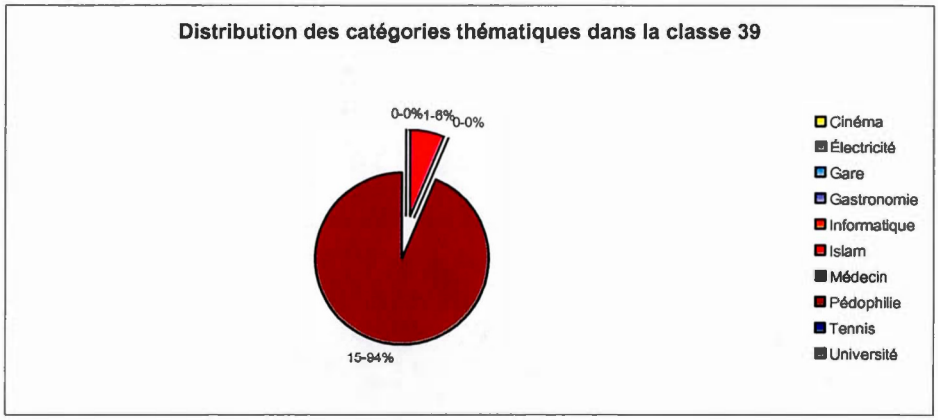

Distribution des catégories thématiques dans la classe 34
0-0%
14-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

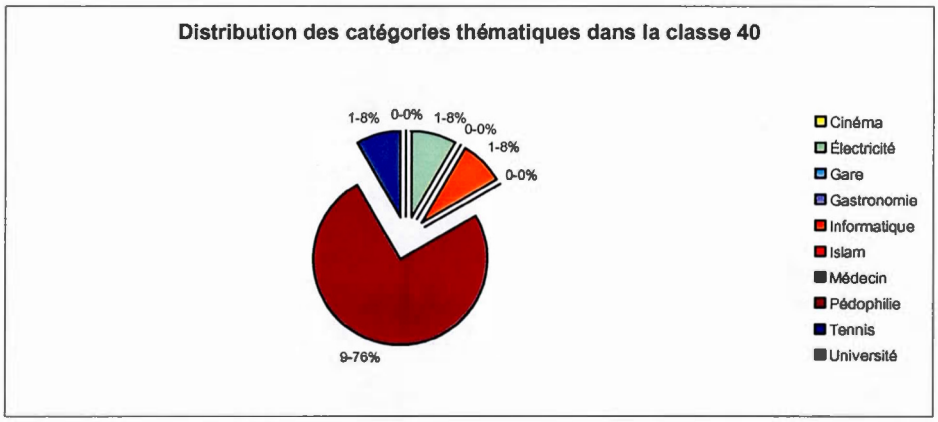

Distribution des catégories thématiques dans la classe 35
0-0%
5-9%
0-0%
3-14%
1-5%
0-0%
12-57%
4-19%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

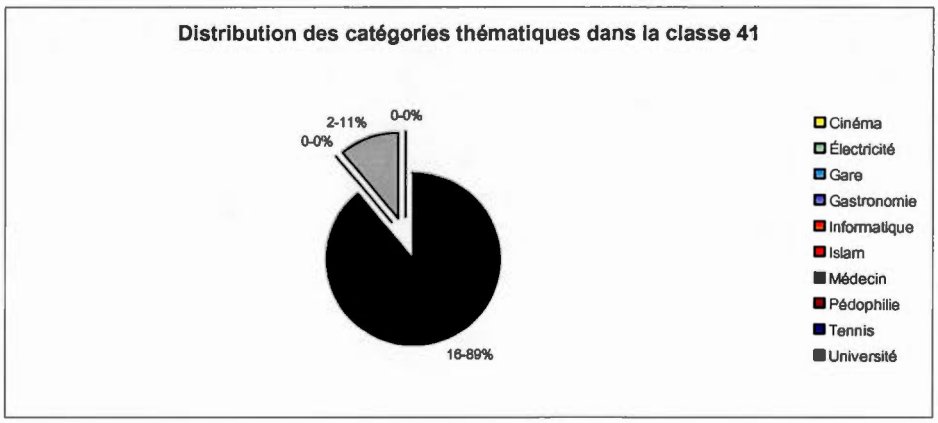

Distribution des catégories thématiques dans la classe 36
0-0%
2-17%
2-17%
0-0%
0-0%
2-17%
3-24%
1-8%
2-17%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

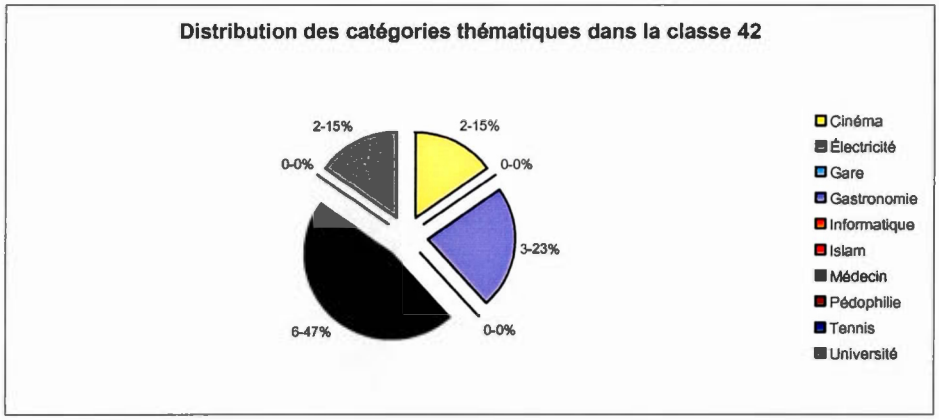

Distribution des catégories thématiques dans la classe 37
36-94%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

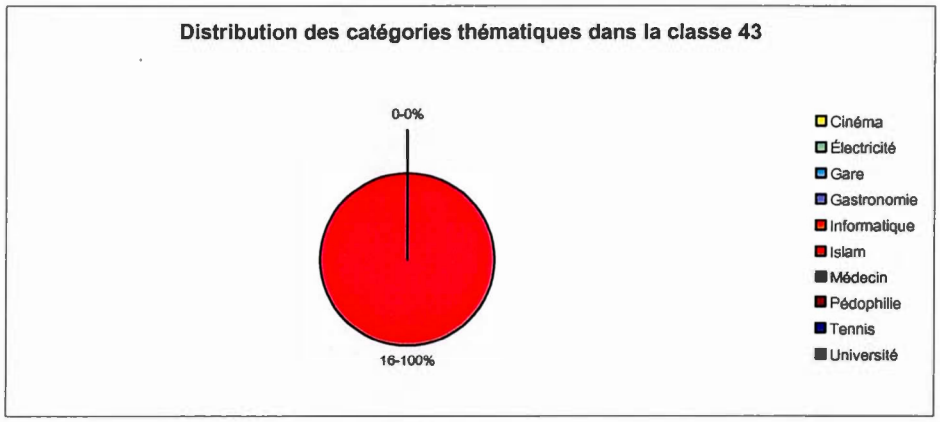

Distribution des catégories thématiques dans la classe 38
1-7%
1-7%
1-7%
0-0%
3-21%
1-7%
1-7%
0-0%
1-7%
5-37%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

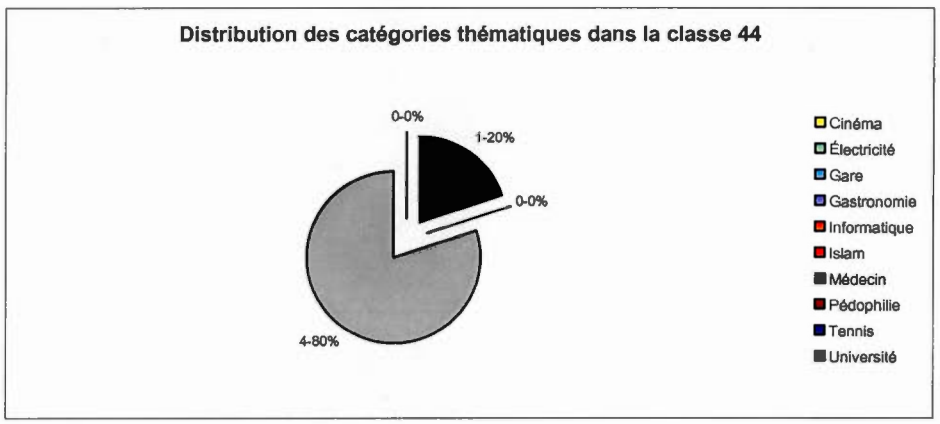

Distribution des catégories thématiques dans la classe 39
0-0%
1-6%
0-0%
15-94%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

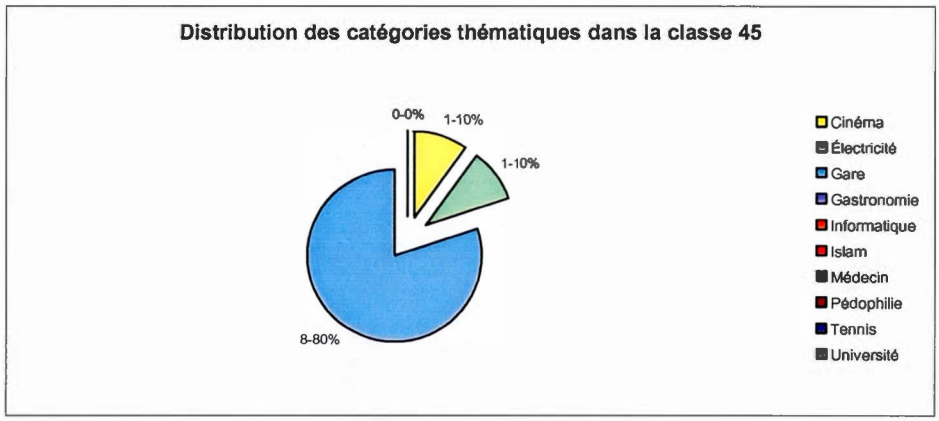

Distribution des catégories thématiques dans la classe 40
1-8%
0-0%
1-8%
0-0%
1-8%
0-0%
9-76%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

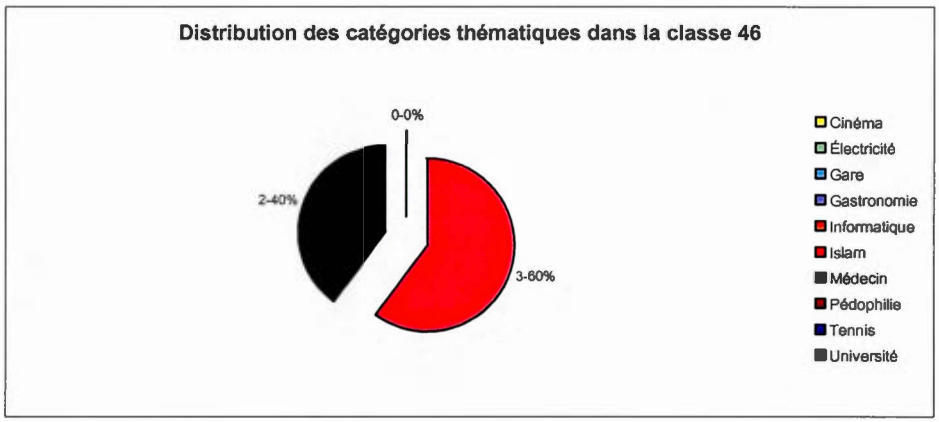

Distribution des catégories thématiques dans la classe 41
2-11%
0-0%
0-0%
16-89%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

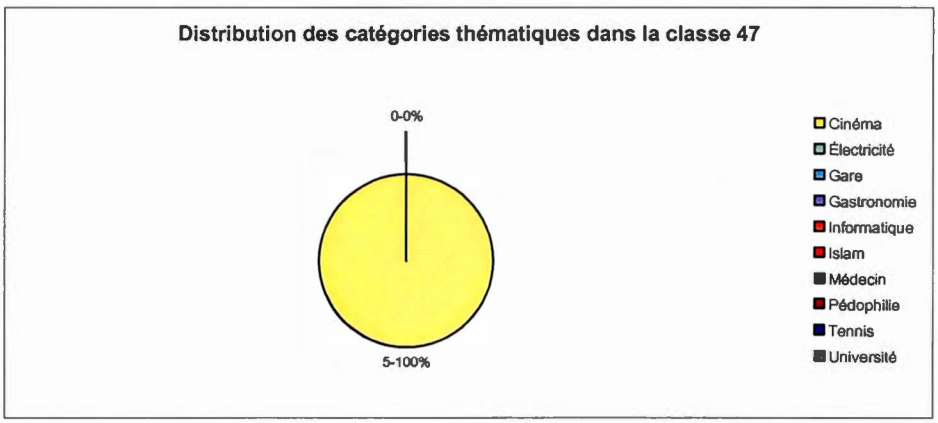

Distribution des catégories thématiques dans la classe 42
2-15%
2-15%
0-0%
0-0%
3-23%
6-47%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 43
0-0%
16-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

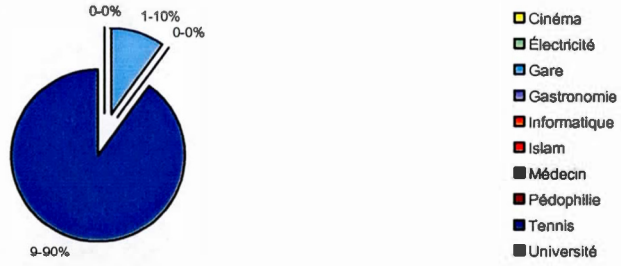

Distribution des catégories thématiques dans la classe 44
0-0%
1-20%
0-0%
4-80%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 45
0-0%
1-10%
1-10%
8-80%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

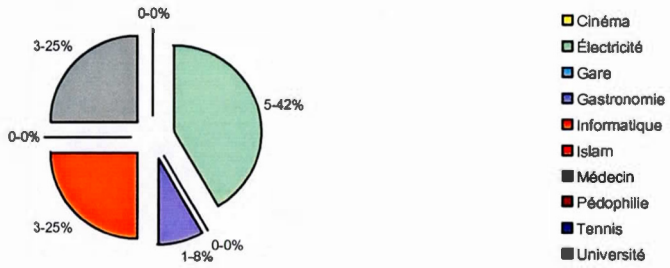

Distribution des catégories thématiques dans la classe 46
0-0%
2-40%
3-60%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 47
0-0%
5-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

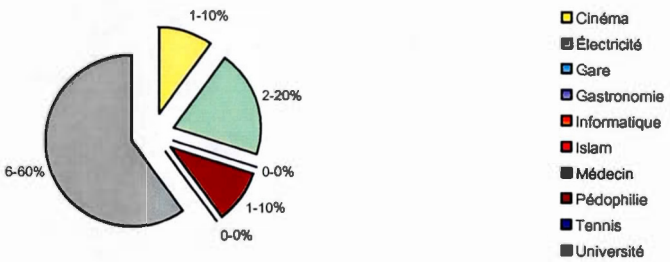
Distribution des catégories thématiques dans la classe 48

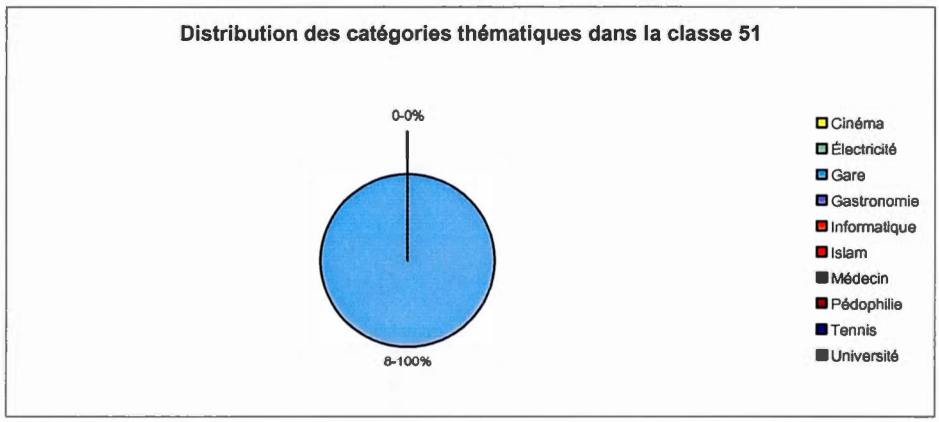


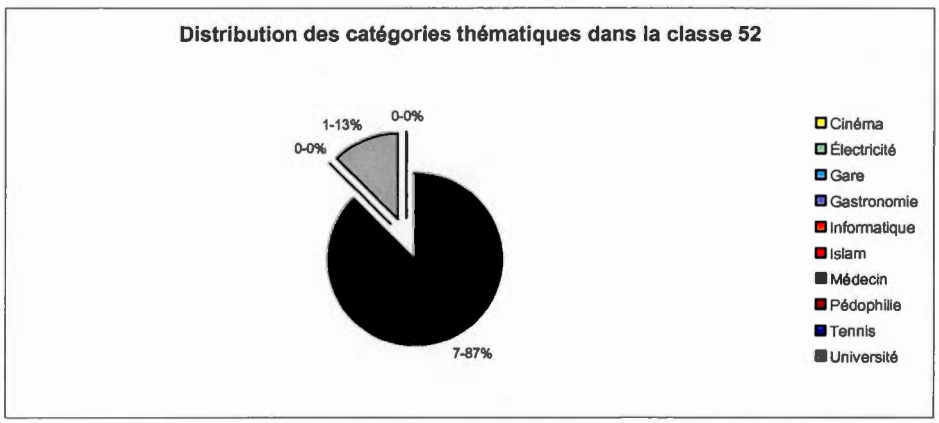
Distribution des catégories thématiques dans la classe 49

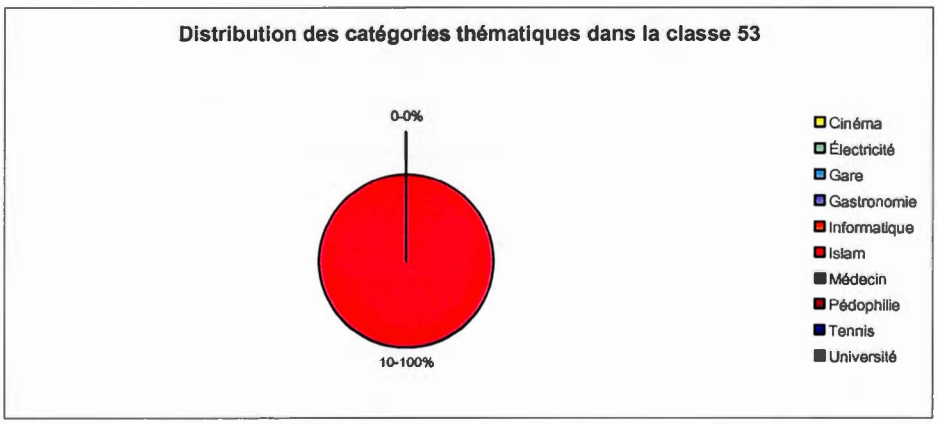


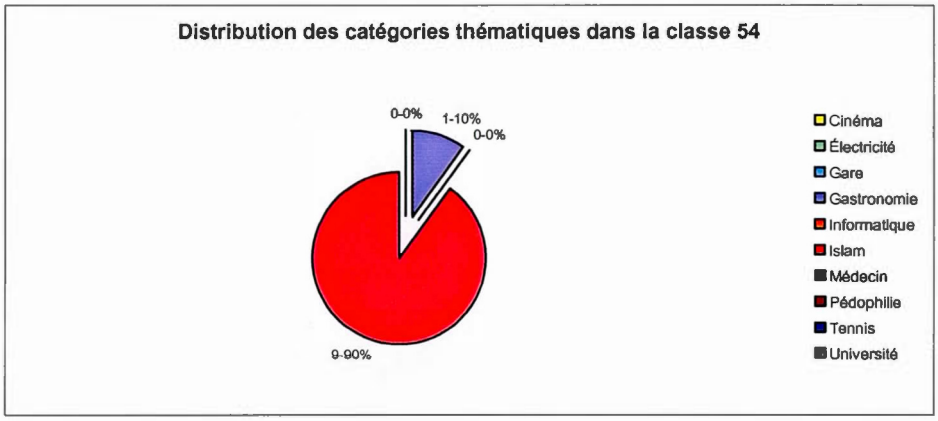
Distribution des catégories thématiques dans la classe 50

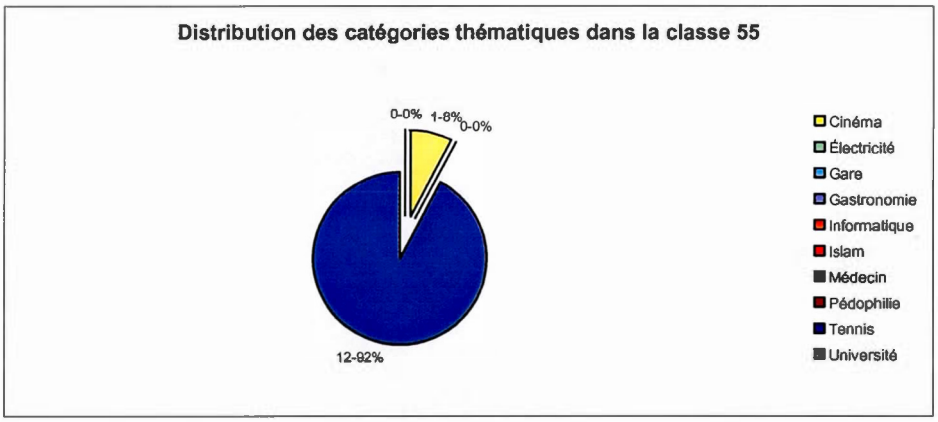


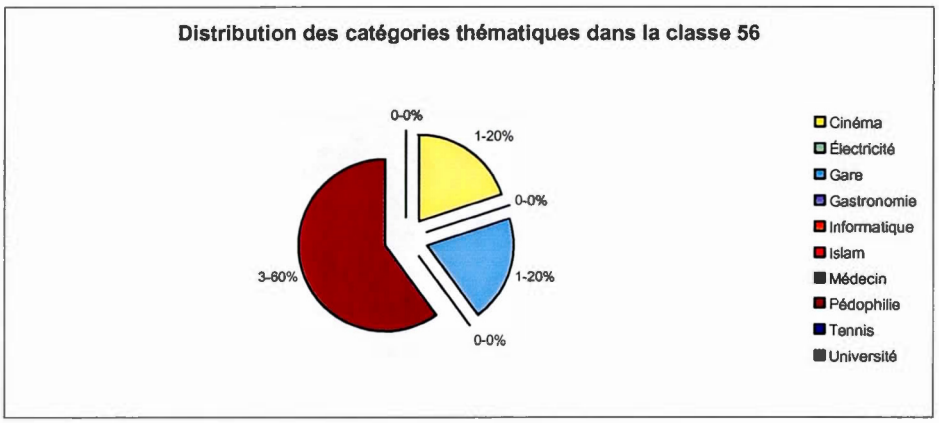

Distribution des catégories thématiques dans la classe 51
0-0%
8-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

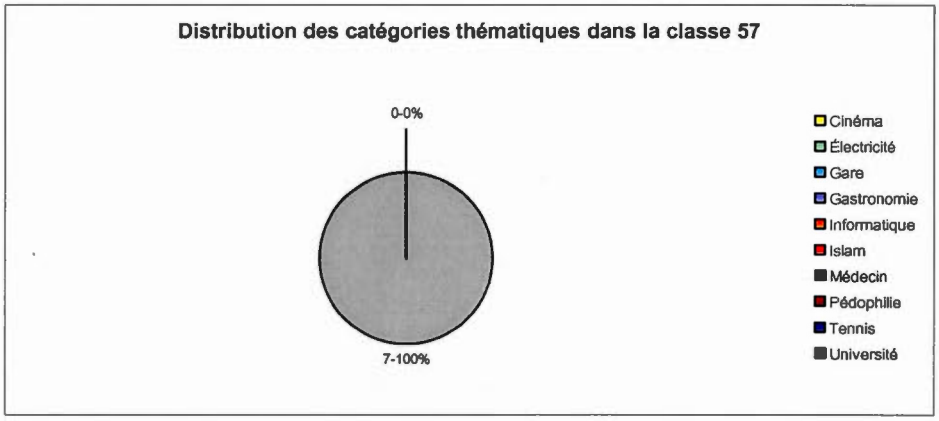

Distribution des catégories thématiques dans la classe 52
1-13%
0-0%
0-0%
7-87%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

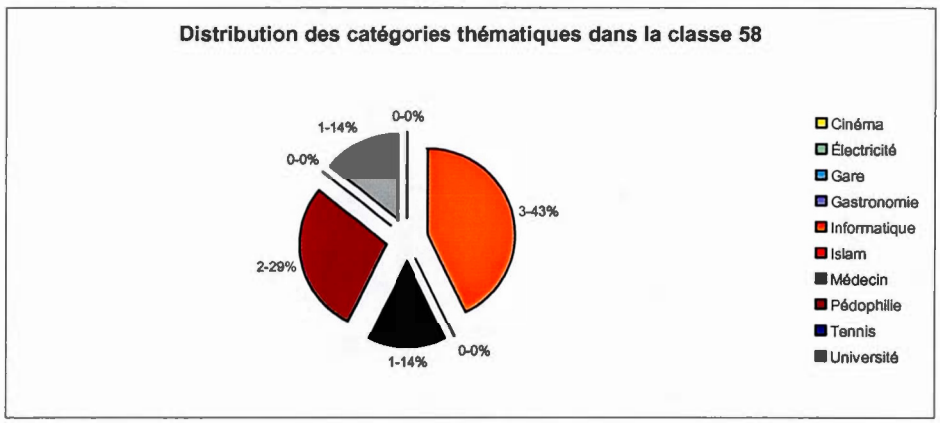

Distribution des catégories thématiques dans la classe 53
0-0%
10-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

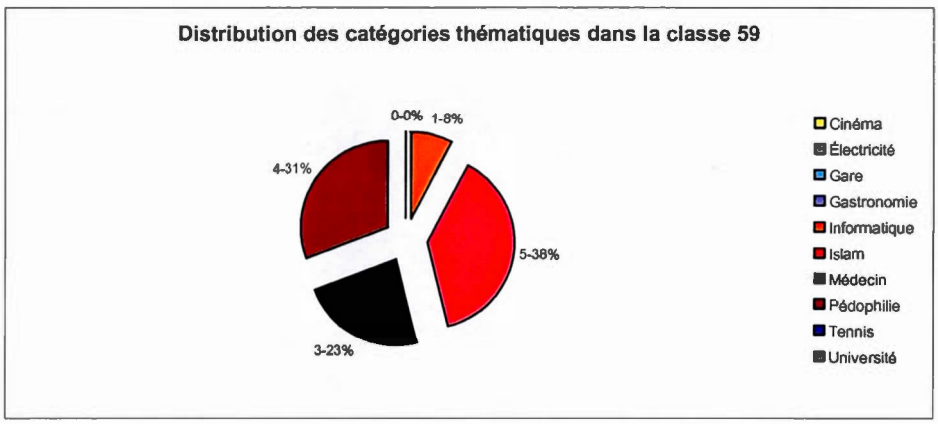

Distribution des catégories thématiques dans la classe 54
0-0%
1-10%
0-0%
9-90%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

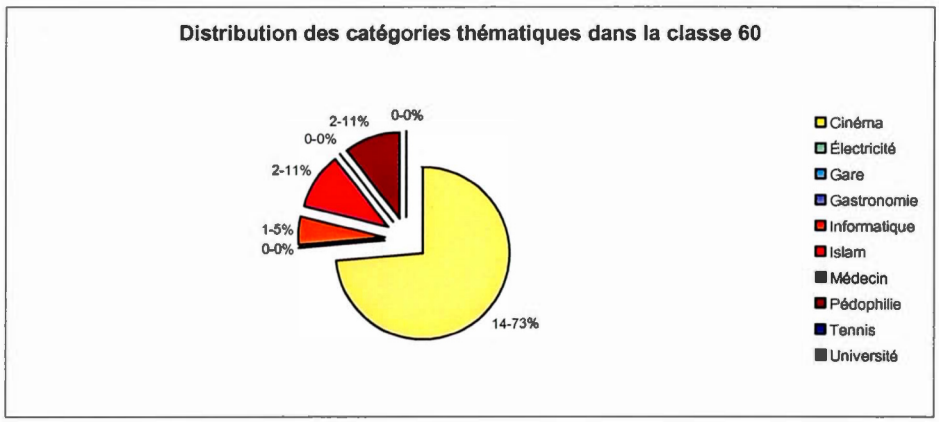

Distribution des catégories thématiques dans la classe 55
0-0%
1-8%
0-0%
12-92%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

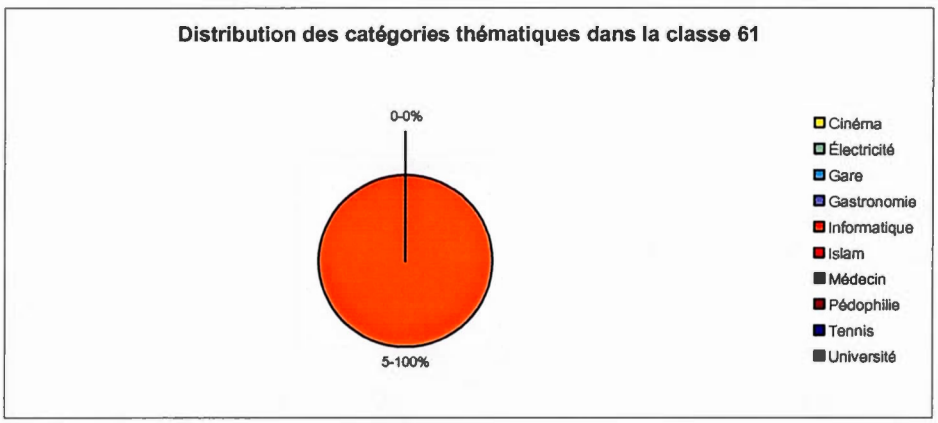

Distribution des catégories thématiques dans la classe 56
0-0%
1-20%
0-0%
1-20%
3-60%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

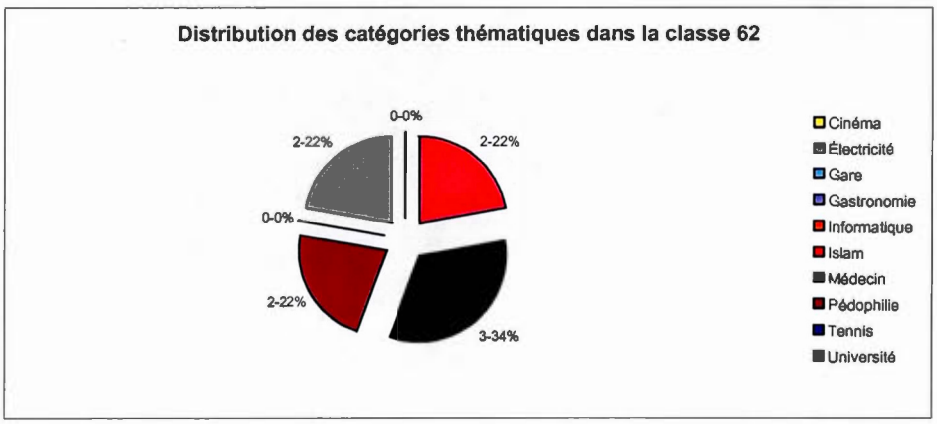

Distribution des catégories thématiques dans la classe 57
0-0%
7-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

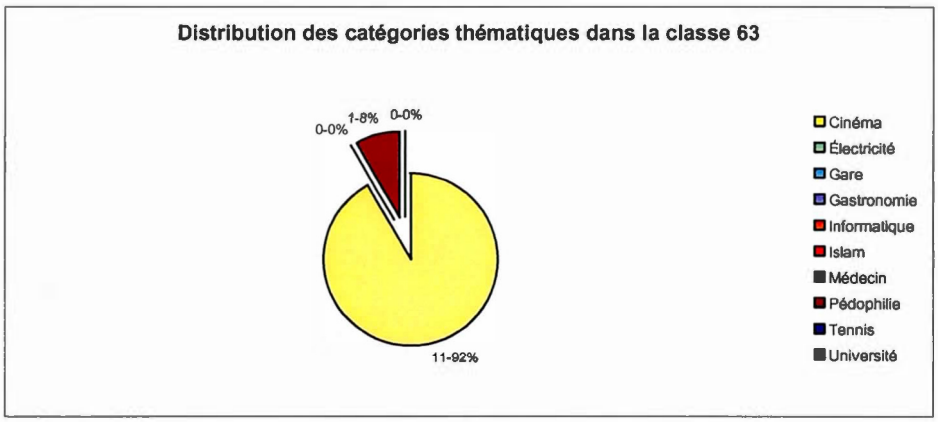

Distribution des catégories thématiques dans la classe 58
1-14%
0-0%
0-0%
3-43%
2-29%
1-14%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

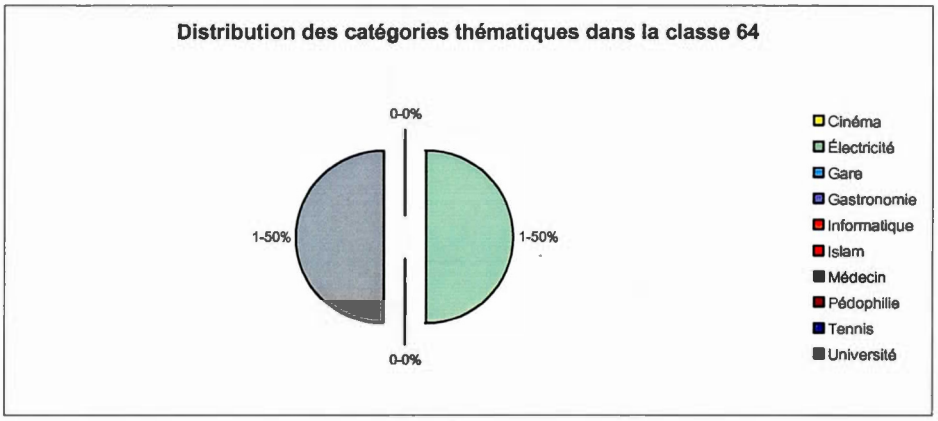

Distribution des catégories thématiques dans la classe 59
0-0%
1-8%
4-31%
5-38%
3-23%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

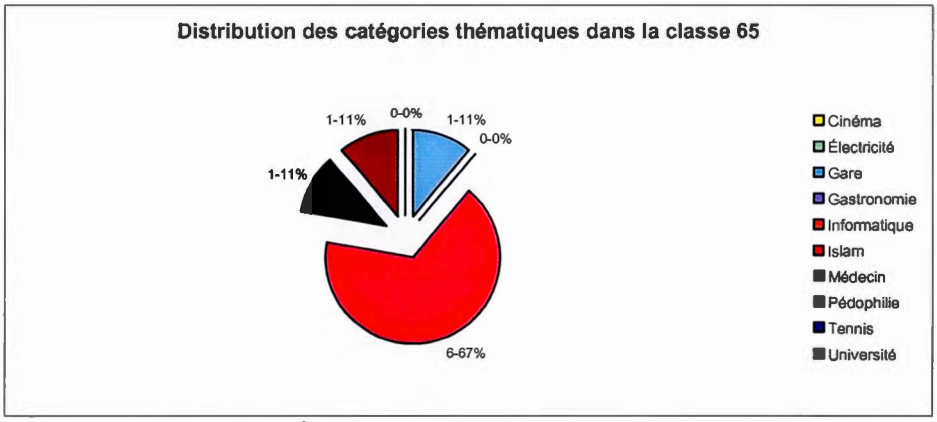

Distribution des catégories thématiques dans la classe 60
2-11%
0-0%
0-0%
2-11%
1-5%
0-0%
14-73%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

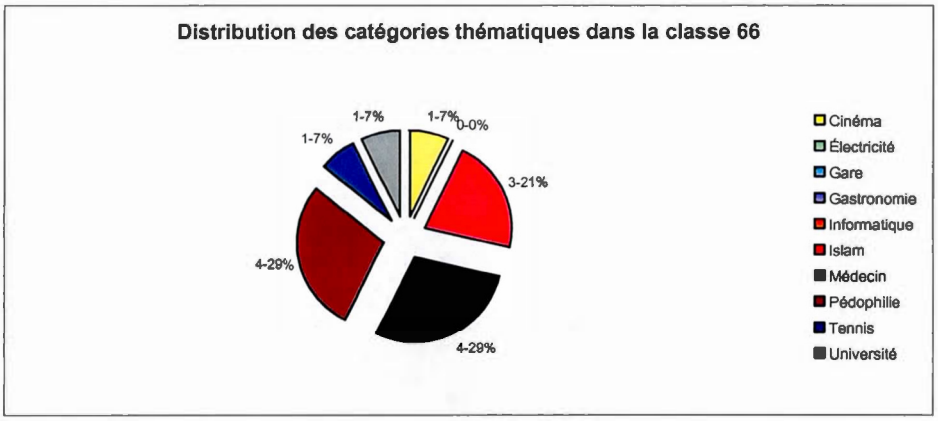

Distribution des catégories thématiques dans la classe 61
0-0%
5-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

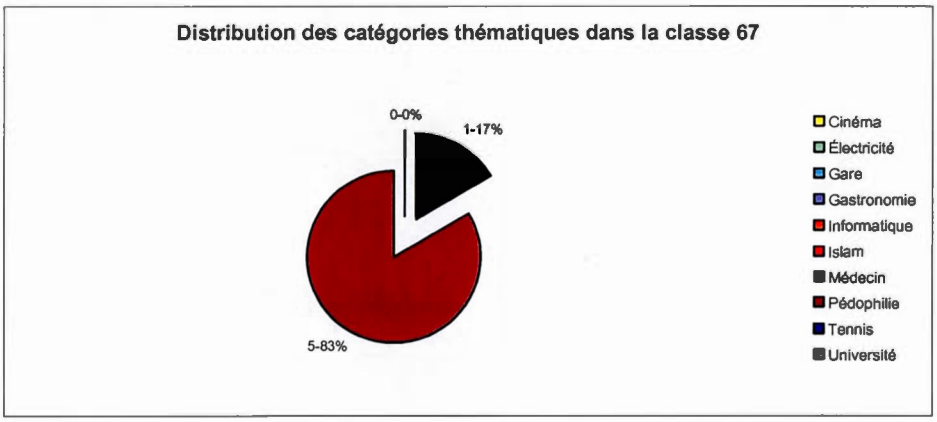

Distribution des catégories thématiques dans la classe 62
0-0%
2-22%
2-22%
0-0%
2-22%
3-34%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

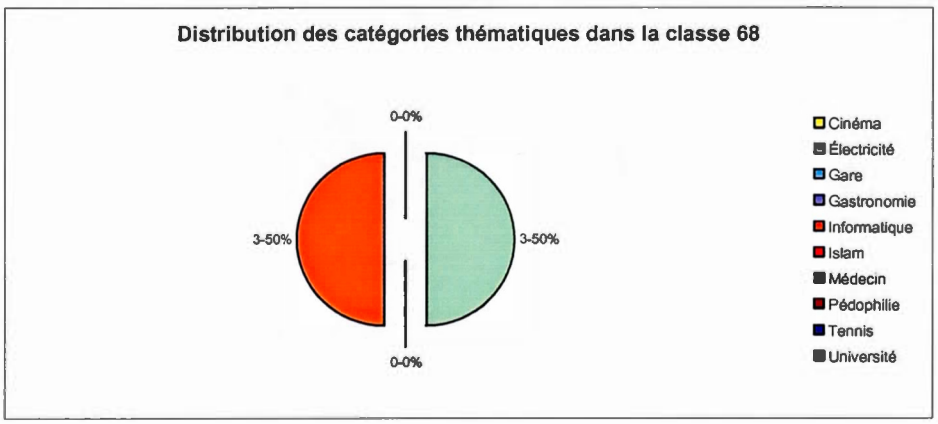

Distribution des catégories thématiques dans la classe 63
0-0%
1-8%
0-0%
11-92%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

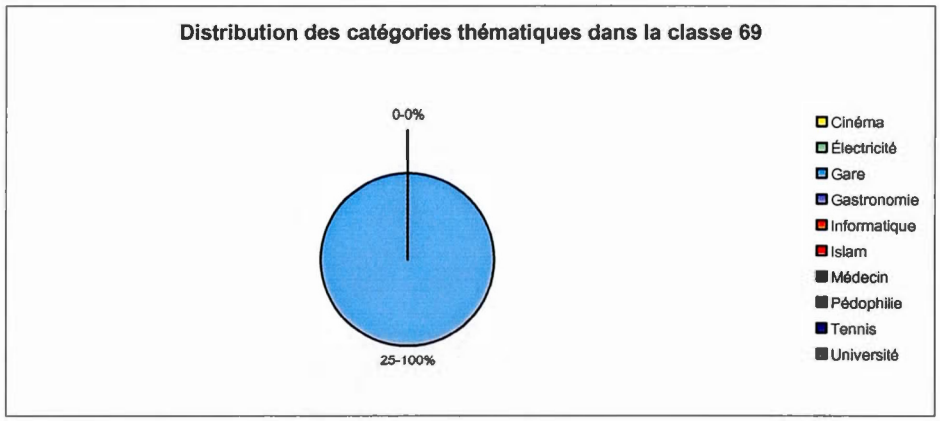

Distribution des catégories thématiques dans la classe 64
0-0%
1-50%
1-50%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

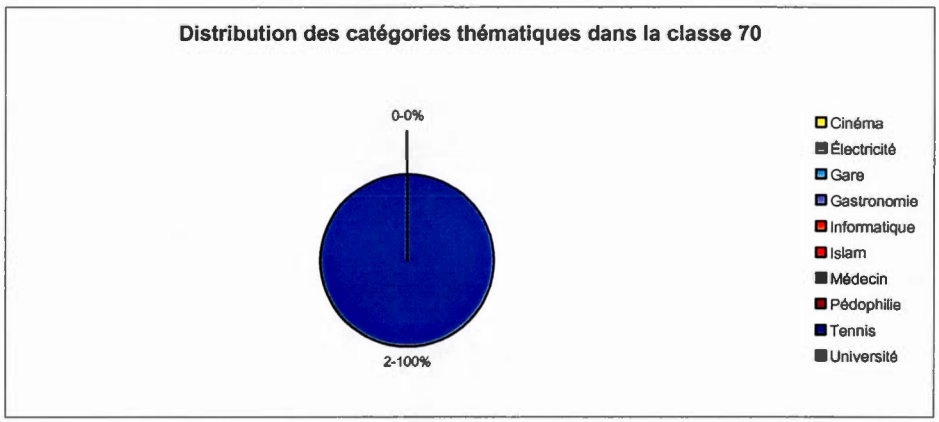

Distribution des catégories thématiques dans la classe 65
1-11%
0-0%
1-11%
0-0%
1-11%
6-67%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

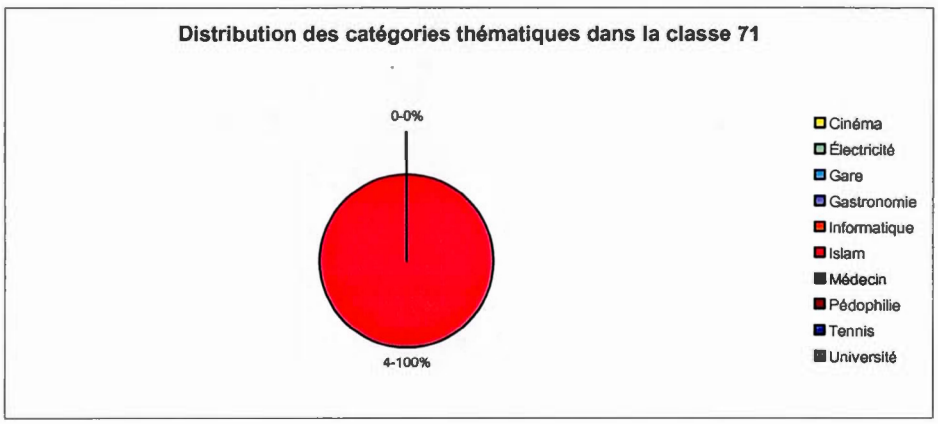

Distribution des catégories thématiques dans la classe 66
1-7%
1-7%
0-0%
1-7%
3-21%
4-29%
4-29%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

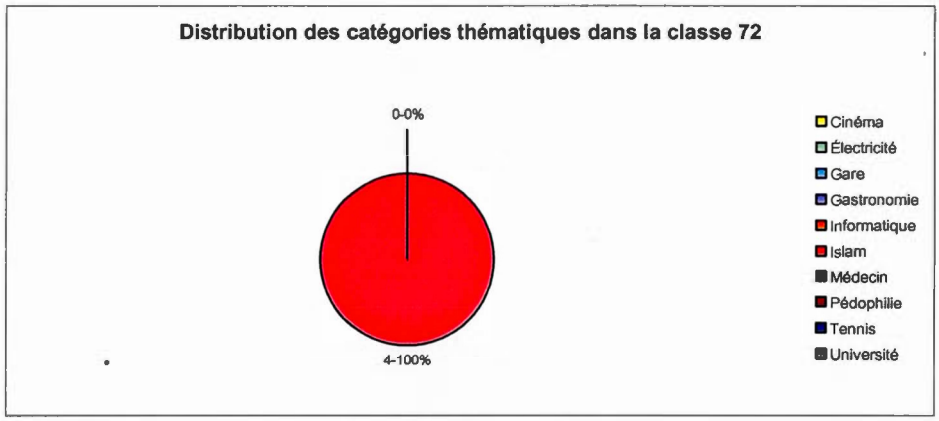

Distribution des catégories thématiques dans la classe 67
0-0%
1-17%
5-83%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

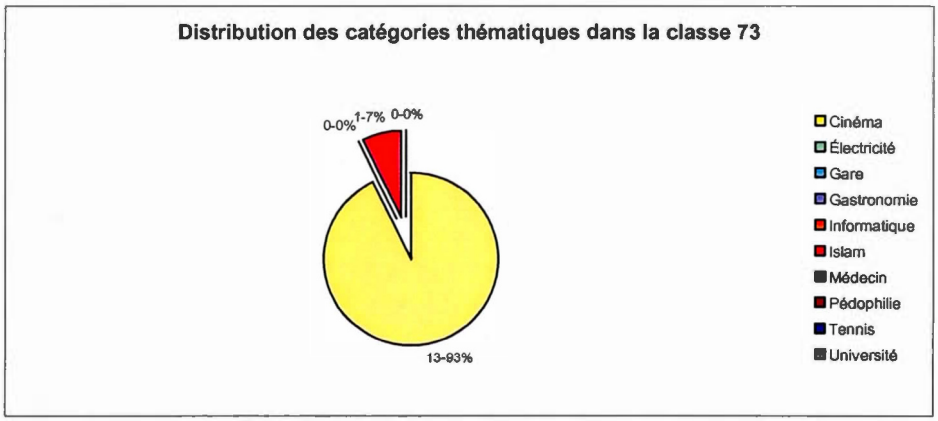

Distribution des catégories thématiques dans la classe 68
0-0%
3-50%
3-50%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

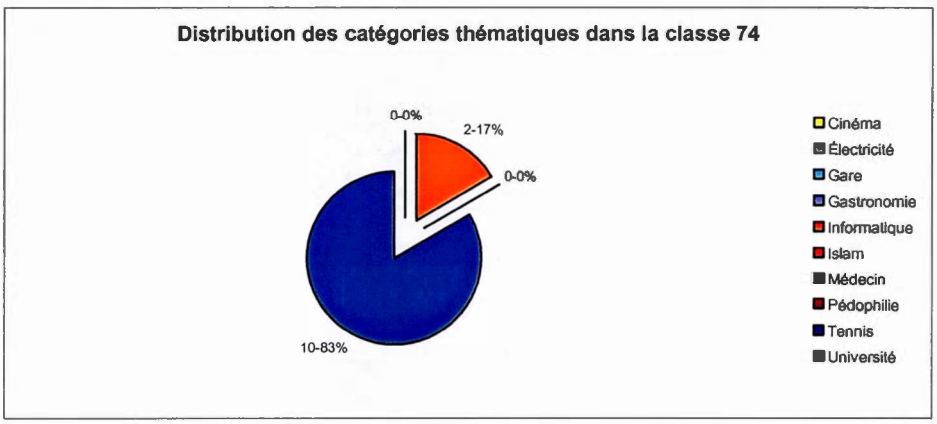

Distribution des catégories thématiques dans la classe 69
0-0%
25-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

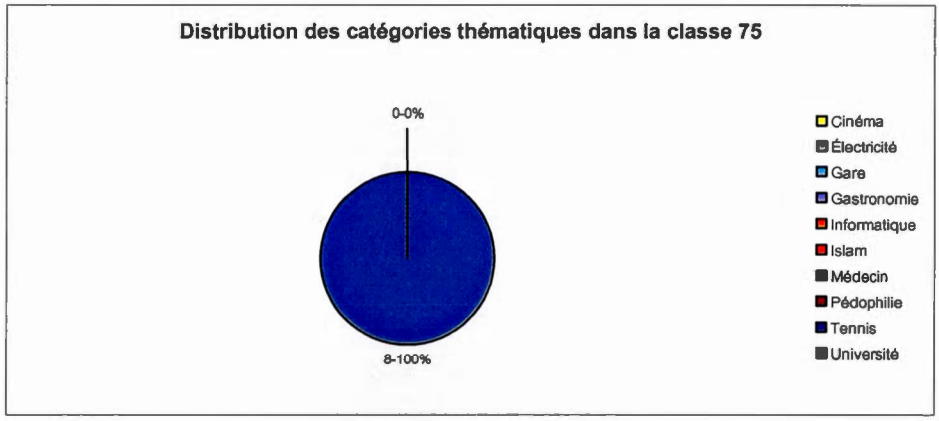

Distribution des catégories thématiques dans la classe 70
0-0%
2-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

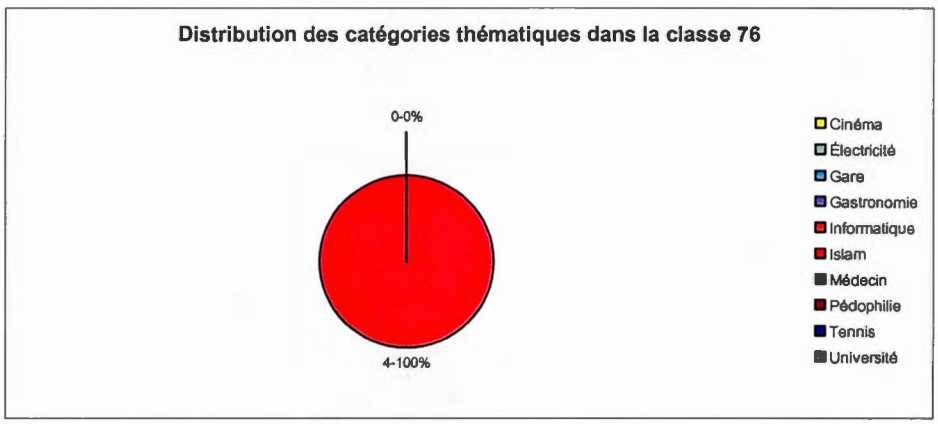

Distribution des catégories thématiques dans la classe 71
0-0%
4-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

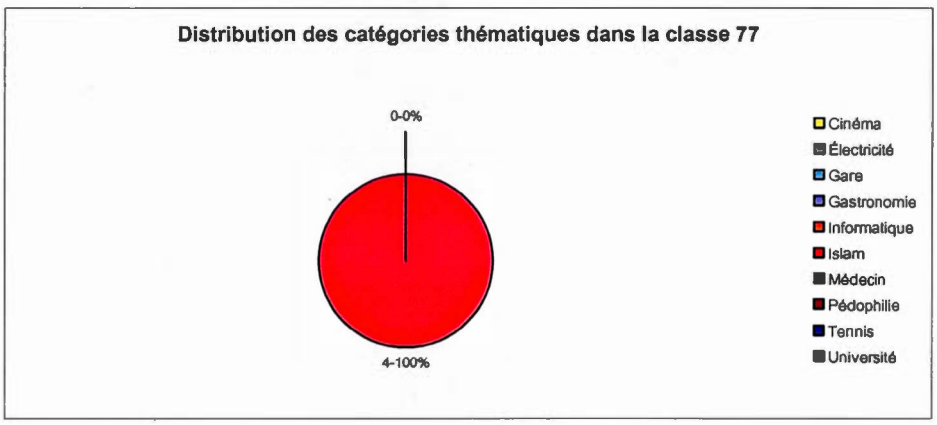

Distribution des catégories thématiques dans la classe 72
0-0%
4-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

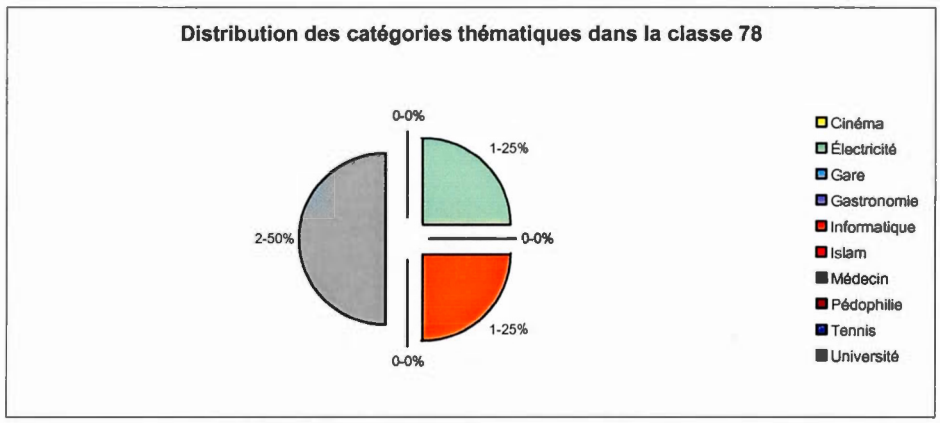

Distribution des catégories thématiques dans la classe 73
0-0%
1-7%
0-0%
13-93%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

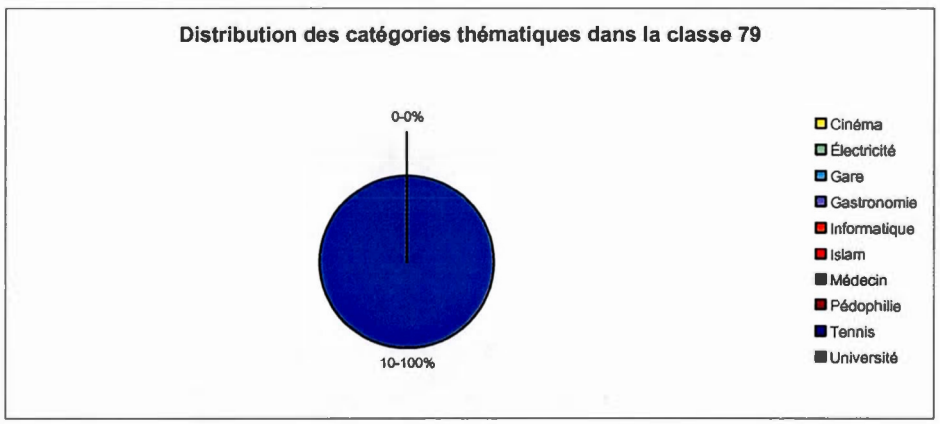

Distribution des catégories thématiques dans la classe 74
0-0%
2-17%
0-0%
10-83%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

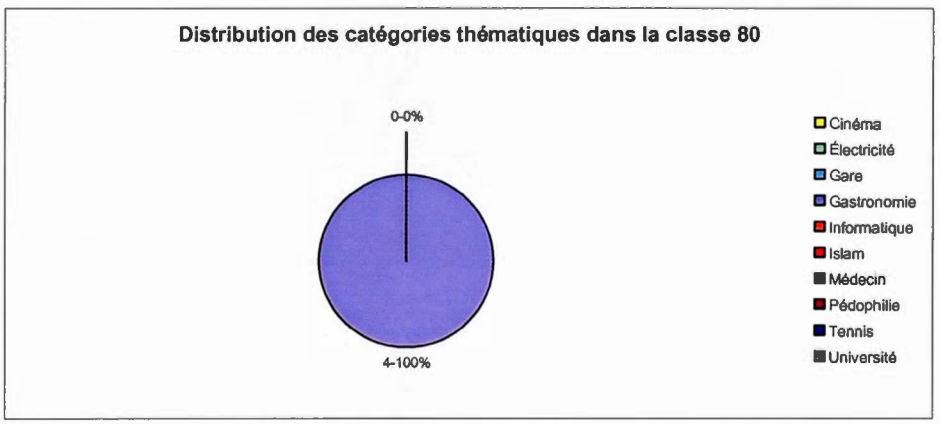

Distribution des catégories thématiques dans la classe 75
0-0%
8-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

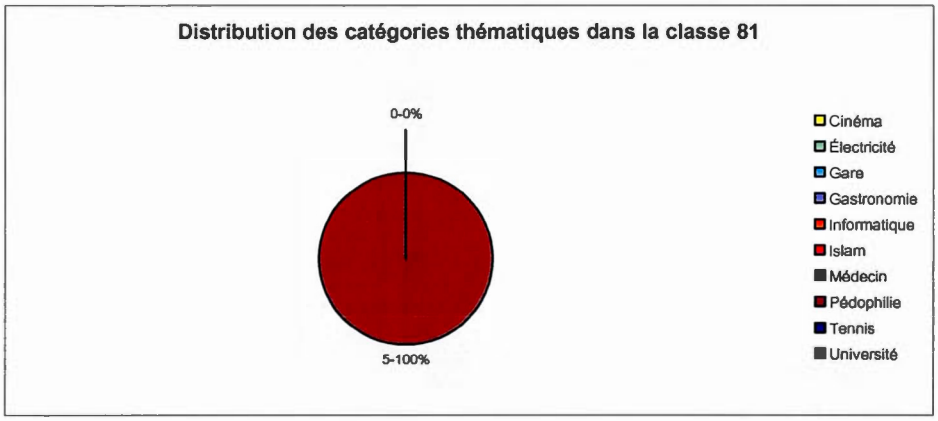

Distribution des catégories thématiques dans la classe 76
0-0%
4-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

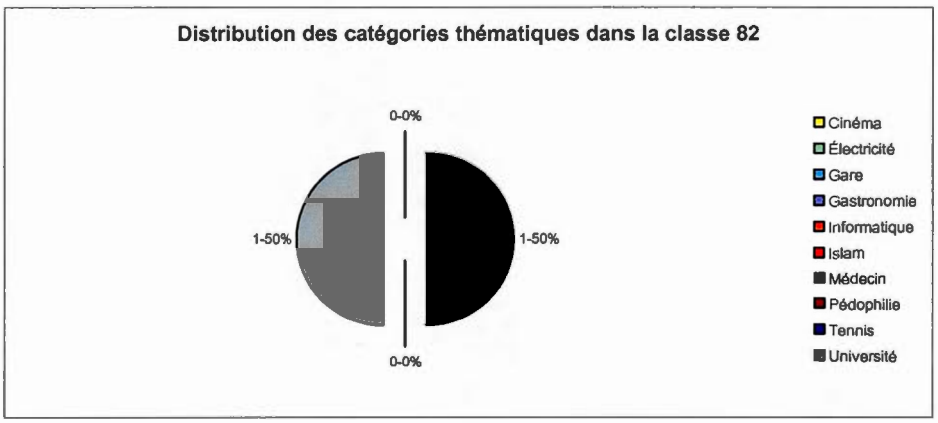

Distribution des catégories thématiques dans la classe 77
0-0%
4-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

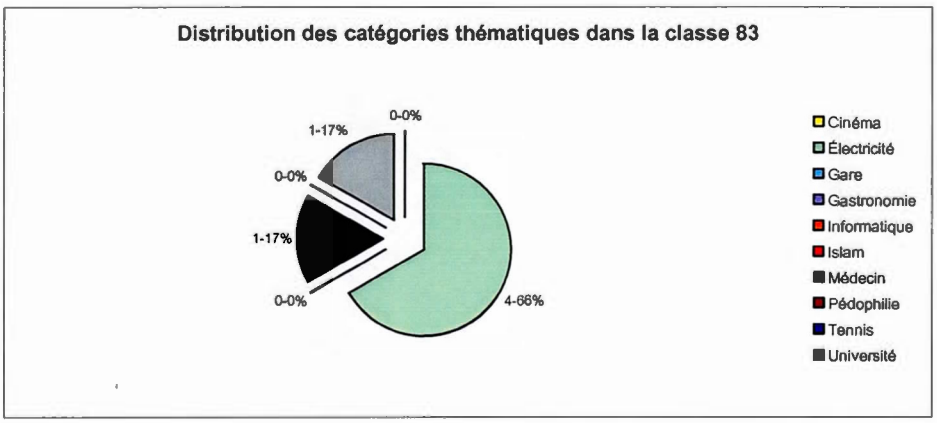

Distribution des catégories thématiques dans la classe 78
0-0%
1-25%
2-50%
0-0%
1-25%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

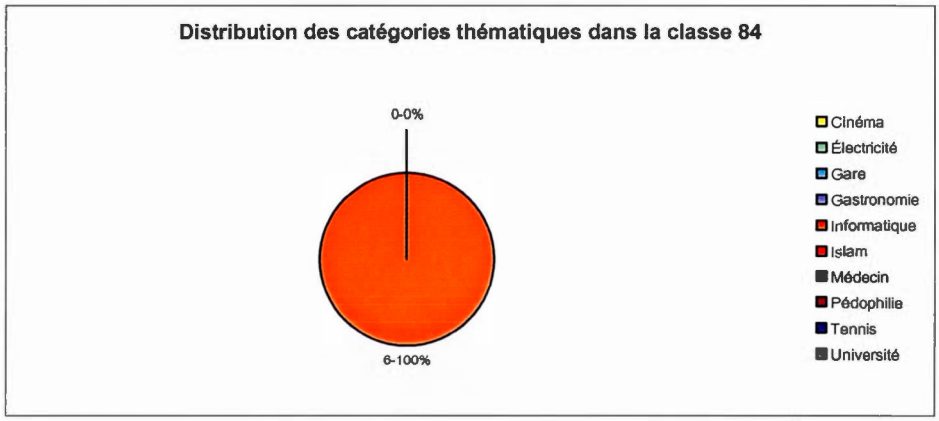

Distribution des catégories thématiques dans la classe 79
0-0%
10-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

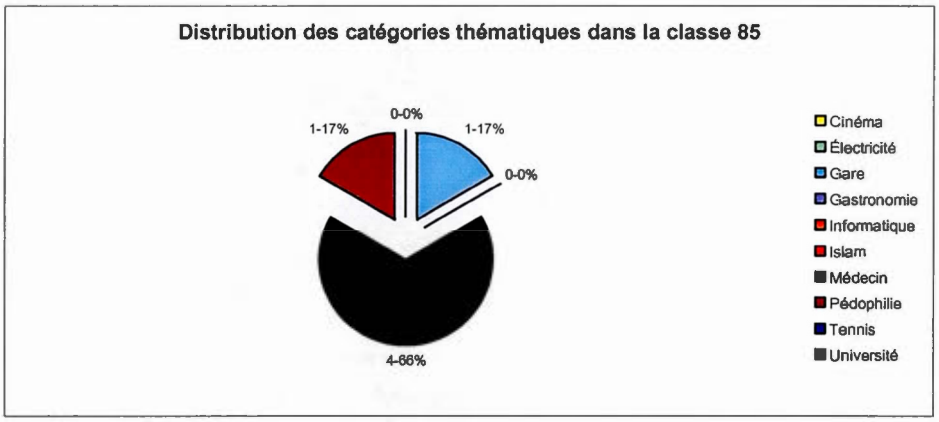

Distribution des catégories thématiques dans la classe 80
0-0%
4-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

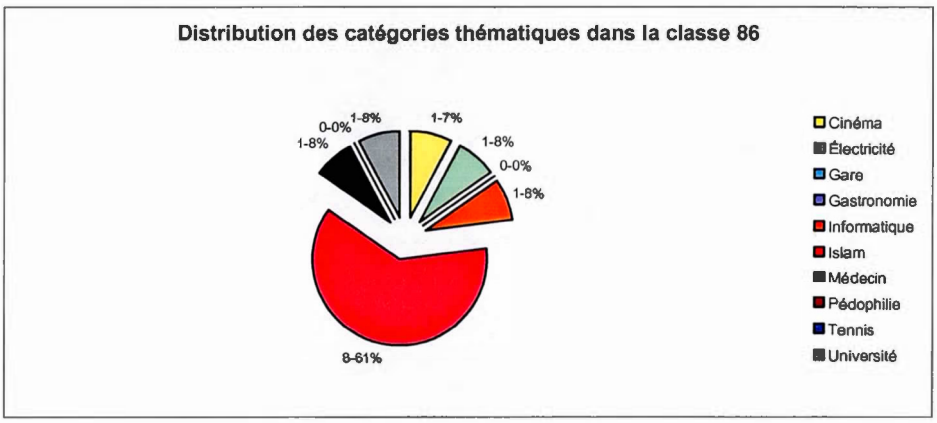

Distribution des catégories thématiques dans la classe 81
0-0%
5-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

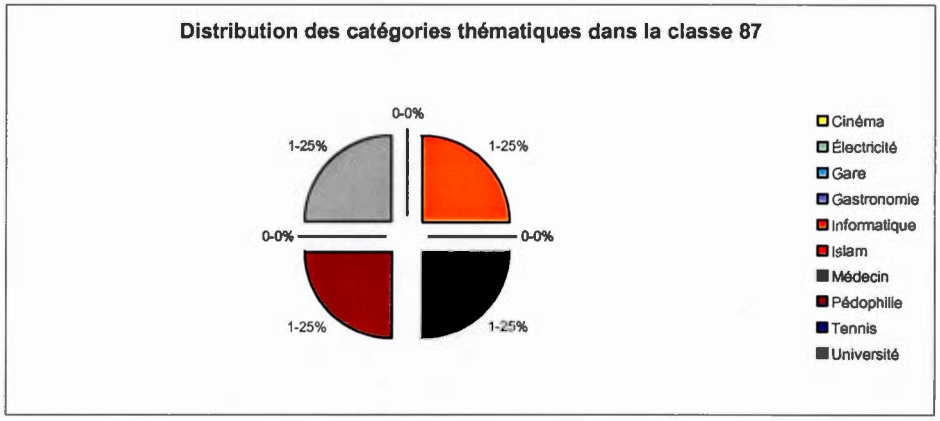

Distribution des catégories thématiques dans la classe 82
0-0%
1-50%
1-50%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

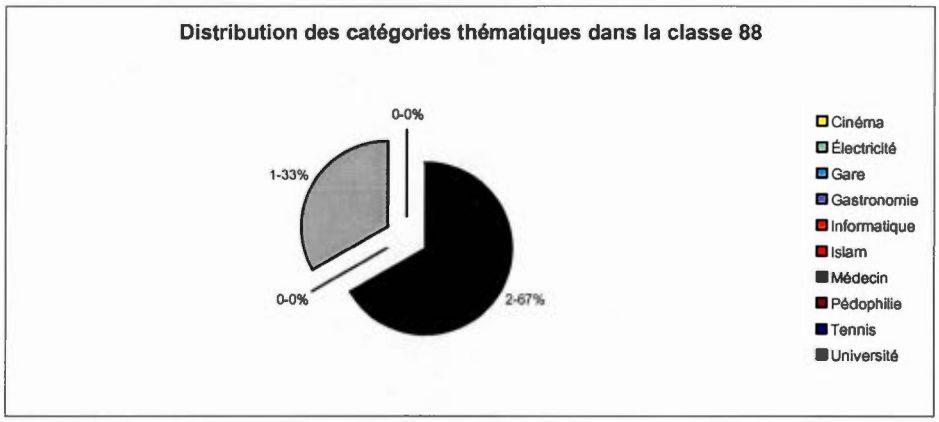

Distribution des catégories thématiques dans la classe 83
0-0%
1-17%
0-0%
1-17%
0-0%
4-66%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

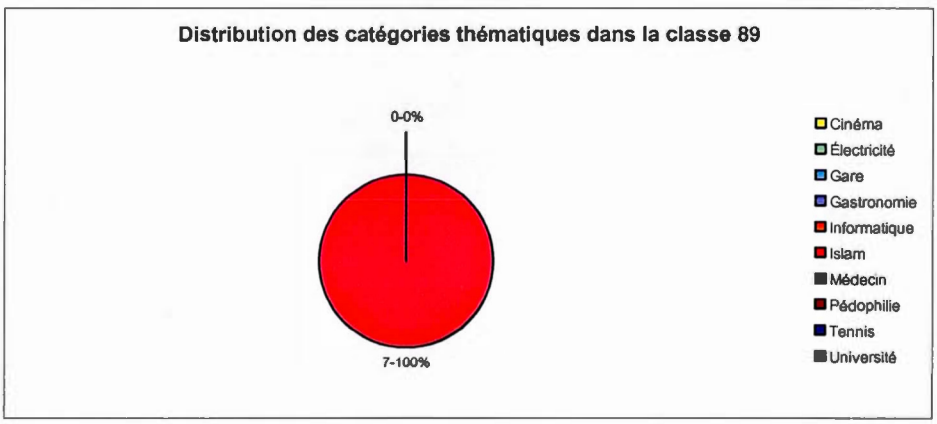

Distribution des catégories thématiques dans la classe 84
0-0%
6-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

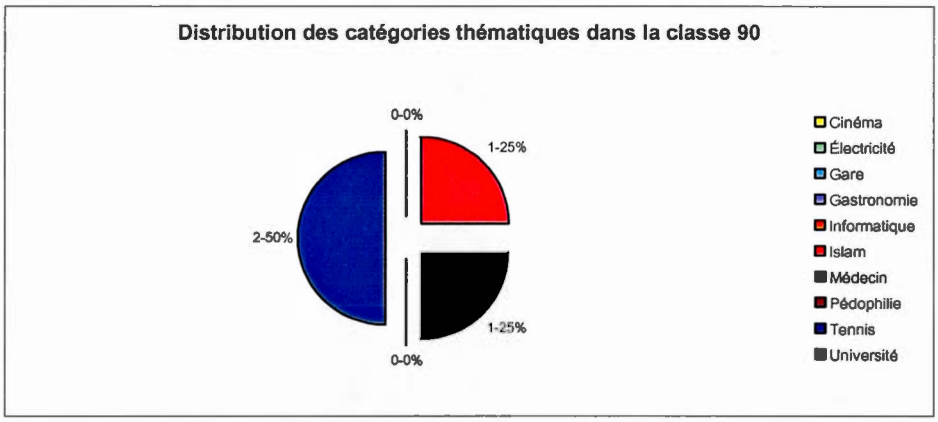

Distribution des catégories thématiques dans la classe 85
0-0%
1-17%
1-17%
0-0%
4-66%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

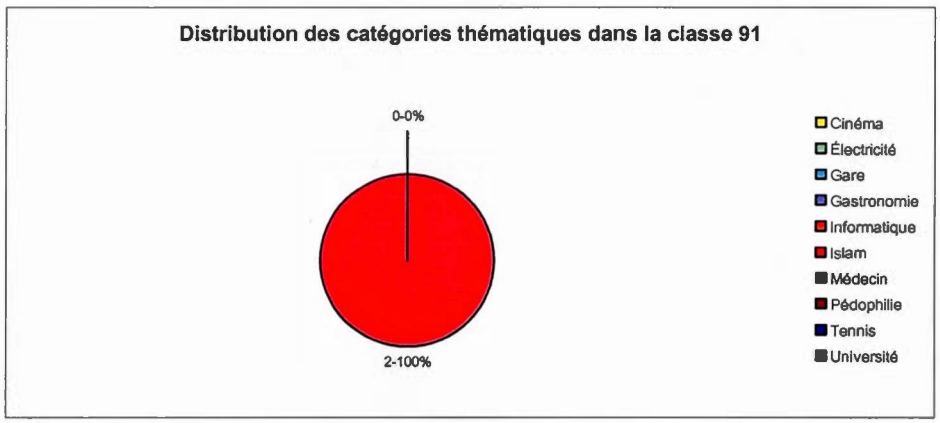

Distribution des catégories thématiques dans la classe 86
0-0%
1-8%
1-7%
1-8%
1-8%
0-0%
1-8%
8-61%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

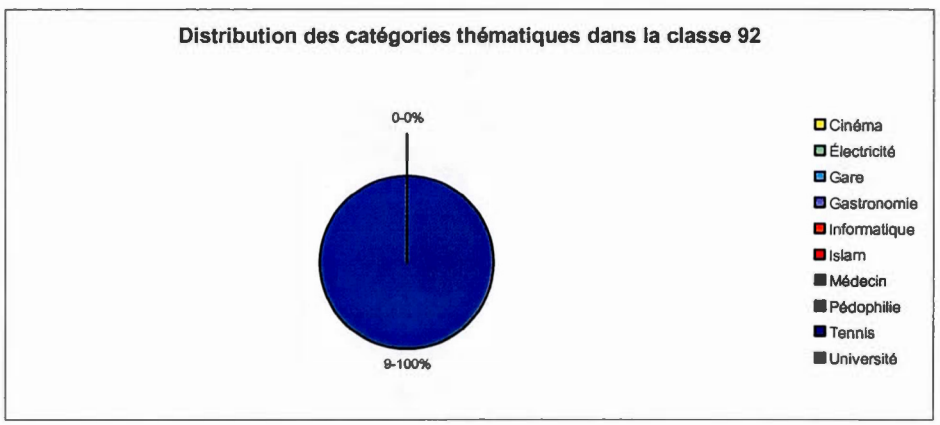

Distribution des catégories thématiques dans la classe 87
0-0%
1-25%
1-25%
0-0%
0-0%
1-25%
1-25%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

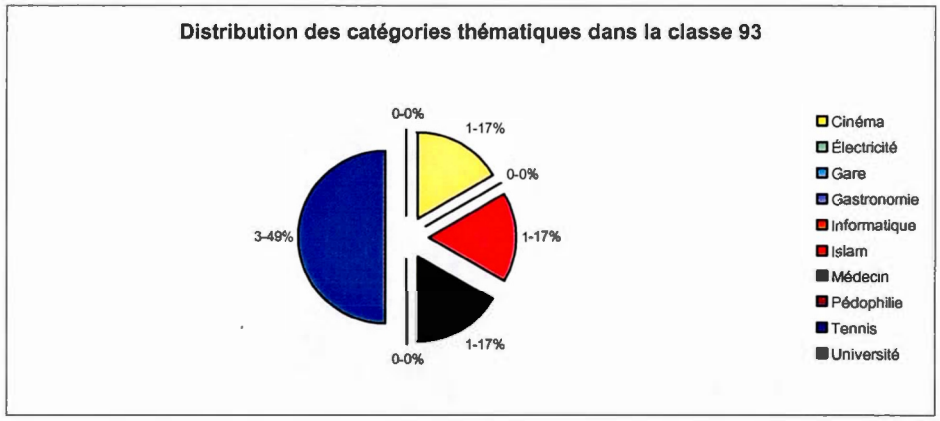

Distribution des catégories thématiques dans la classe 88
0-0%
1-33%
0-0%
2-67%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

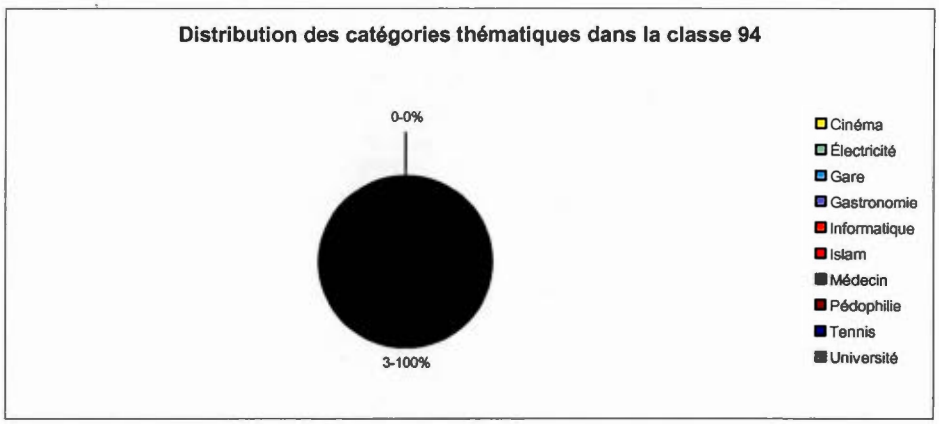

Distribution des catégories thématiques dans la classe 89
0-0%
7-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

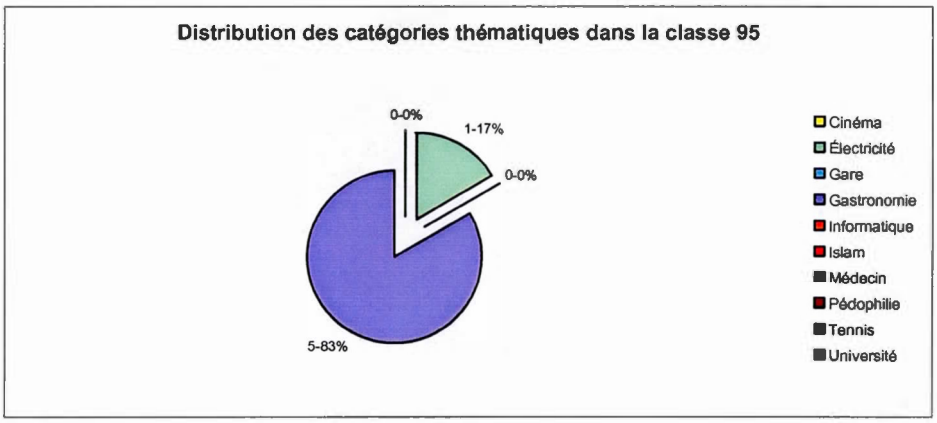

Distribution des catégories thématiques dans la classe 90
0-0%
1-25%
2-50%
1-25%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

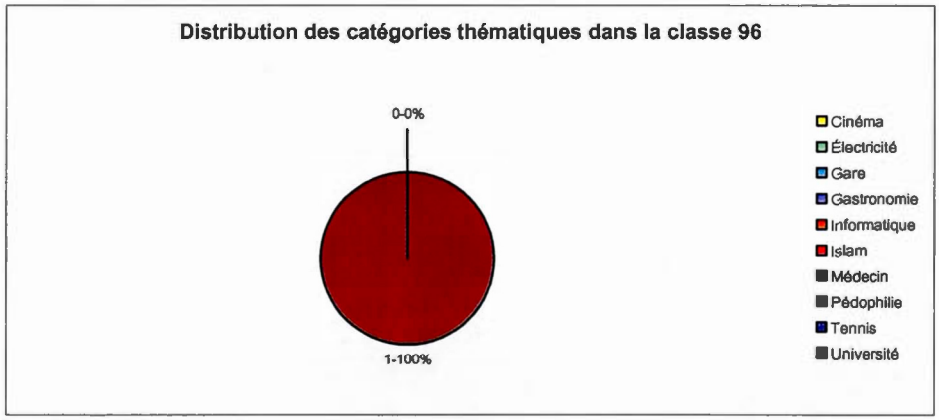

Distribution des catégories thématiques dans la classe 91
0-0%
2-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

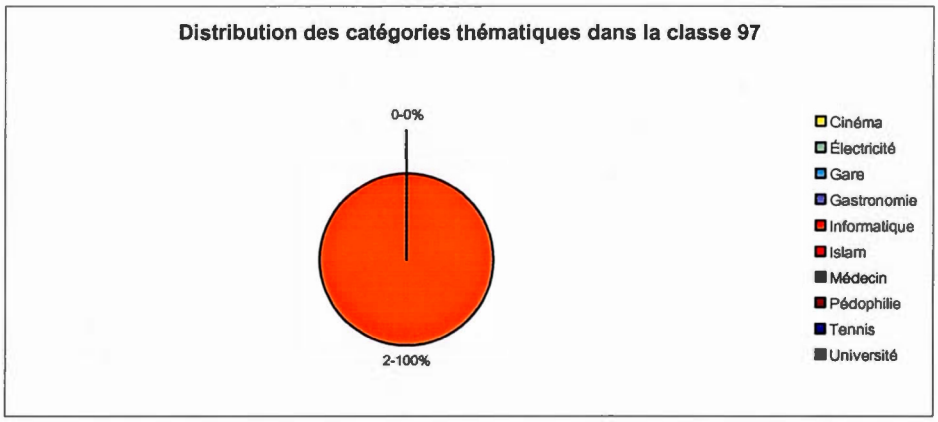

Distribution des catégories thématiques dans la classe 92
0-0%
9-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

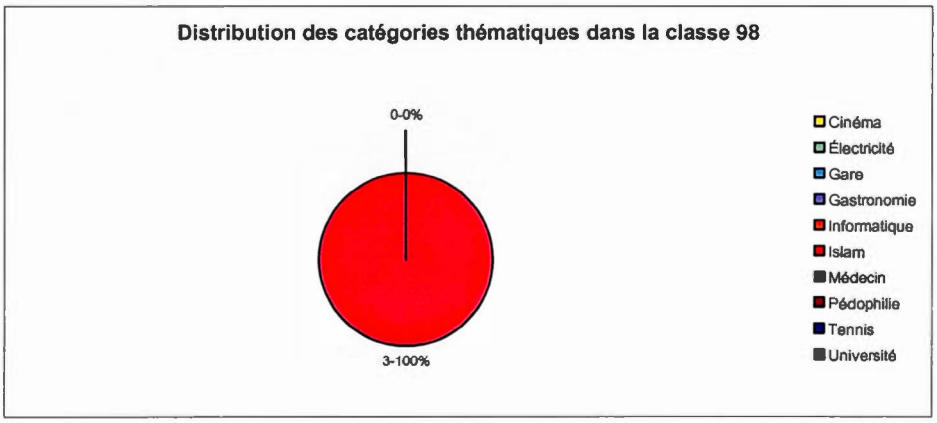

Distribution des catégories thématiques dans la classe 93
0-0%
1-17%
0-0%
1-17%
3-49%
1-17%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

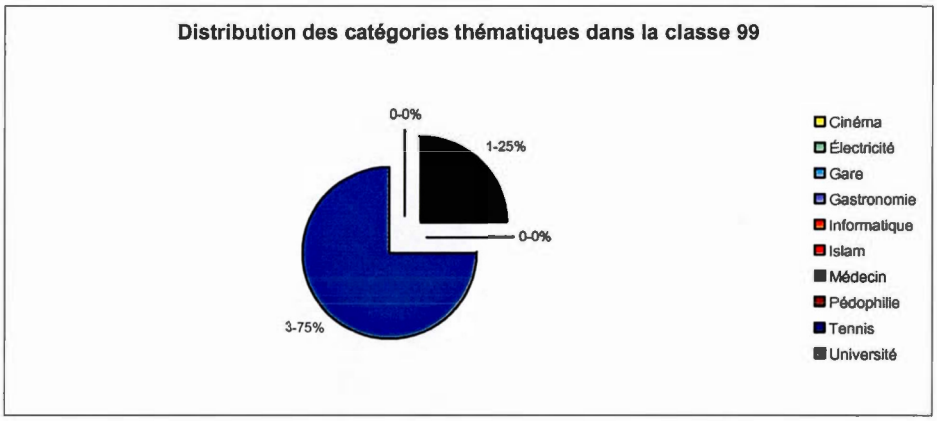

Distribution des catégories thématiques dans la classe 94
0-0%
3-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

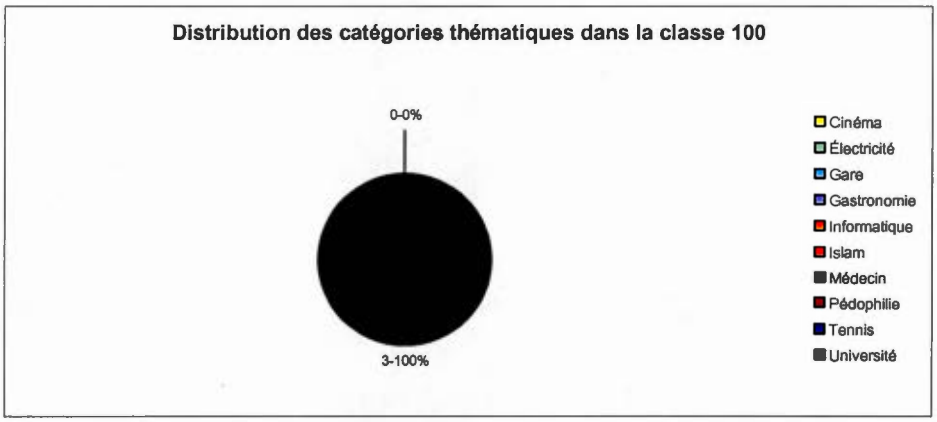

Distribution des catégories thématiques dans la classe 95
0-0%
1-17%
0-0%
5-83%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

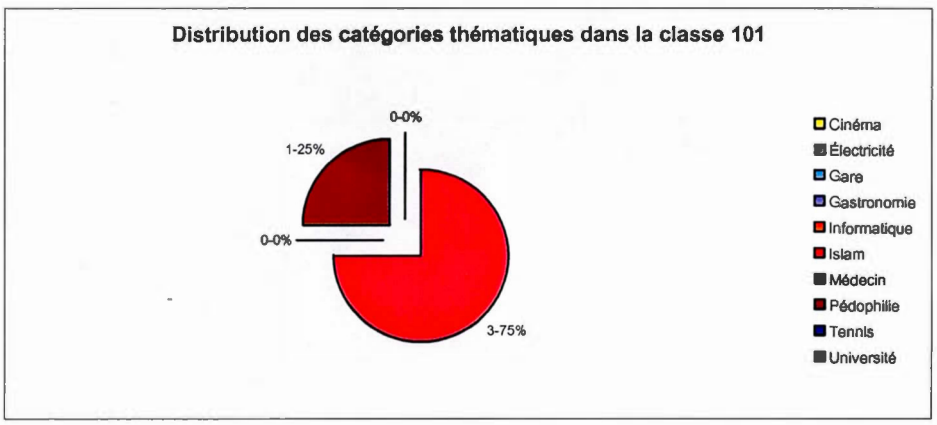

Distribution des catégories thématiques dans la classe 96
0-0%
1-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

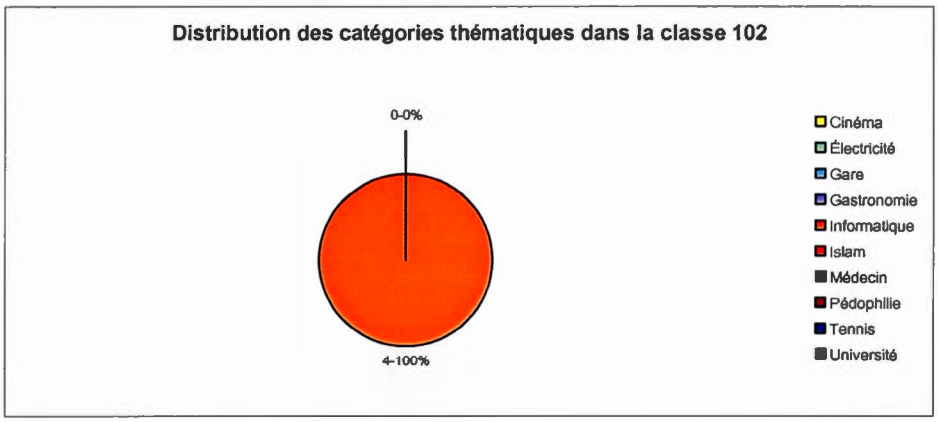

Distribution des catégories thématiques dans la classe 97
0-0%
2-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

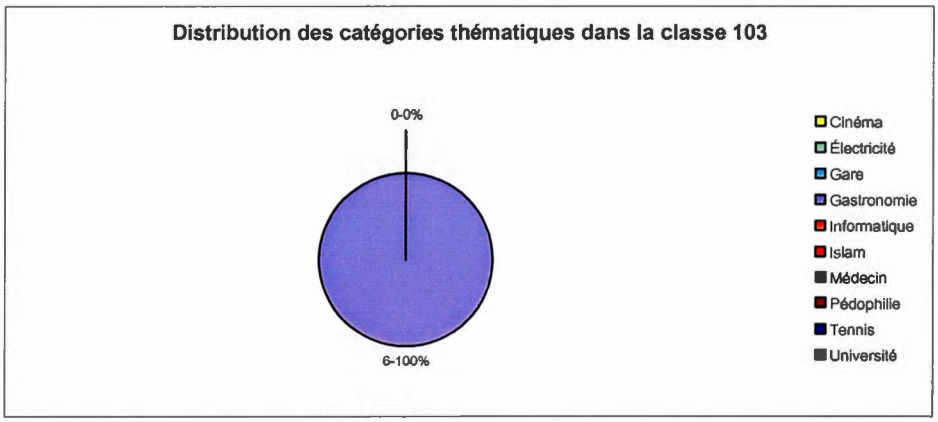

Distribution des catégories thématiques dans la classe 98
0-0%
3-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

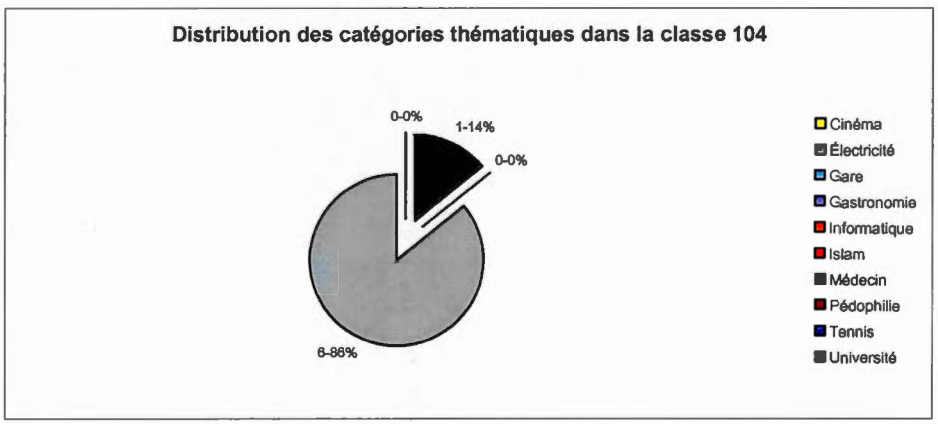

Distribution des catégories thématiques dans la classe 99
0-0%
1-25%
0-0%
3-75%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

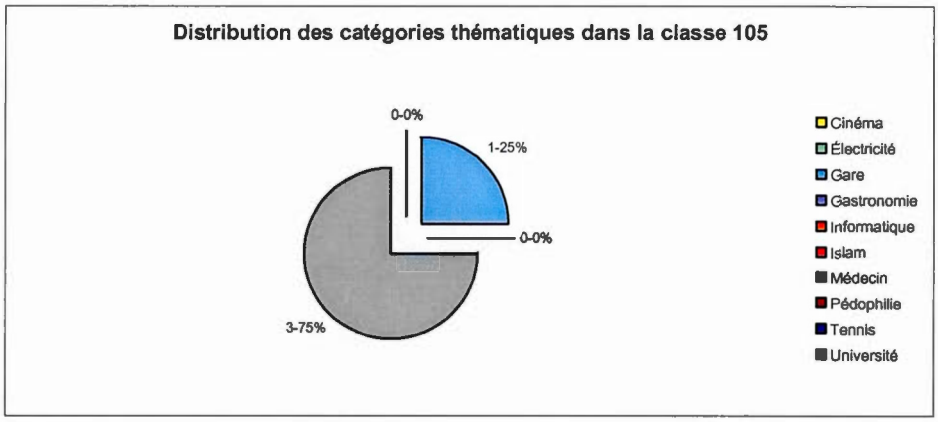

Distribution des catégories thématiques dans la classe 100
0-0%
3-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

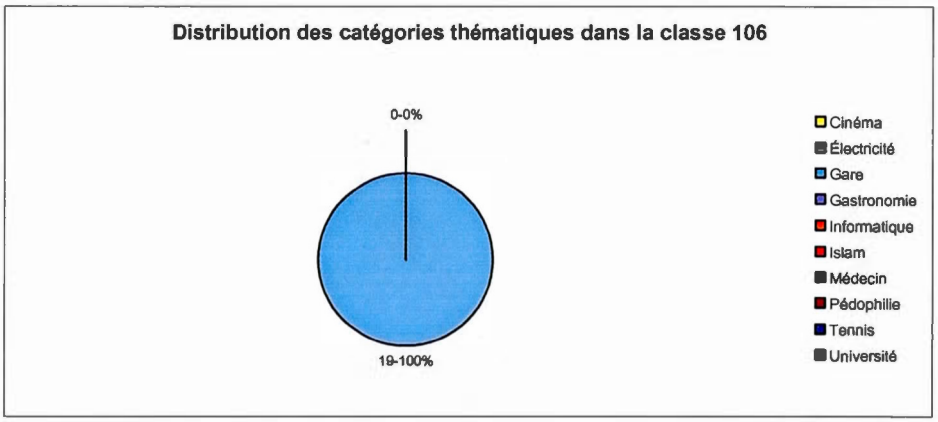

Distribution des catégories thématiques dans la classe 101
0-0%
1-25%
0-0%
3-75%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

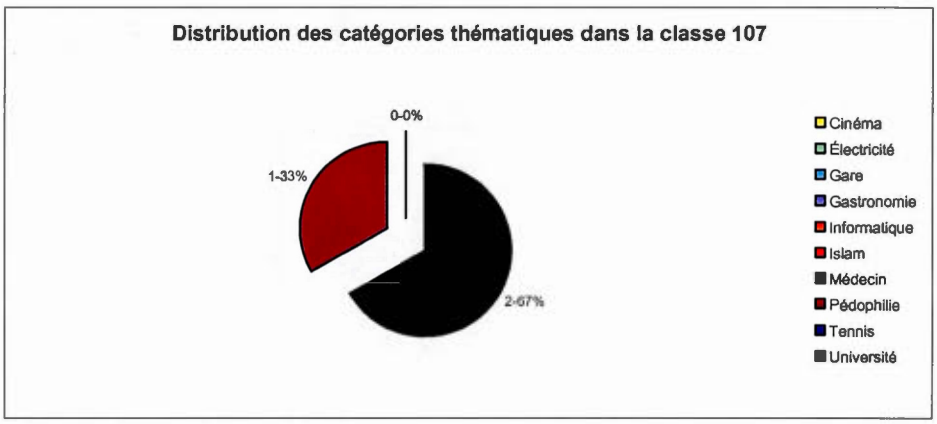

Distribution des catégories thématiques dans la classe 102
0-0%
4-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

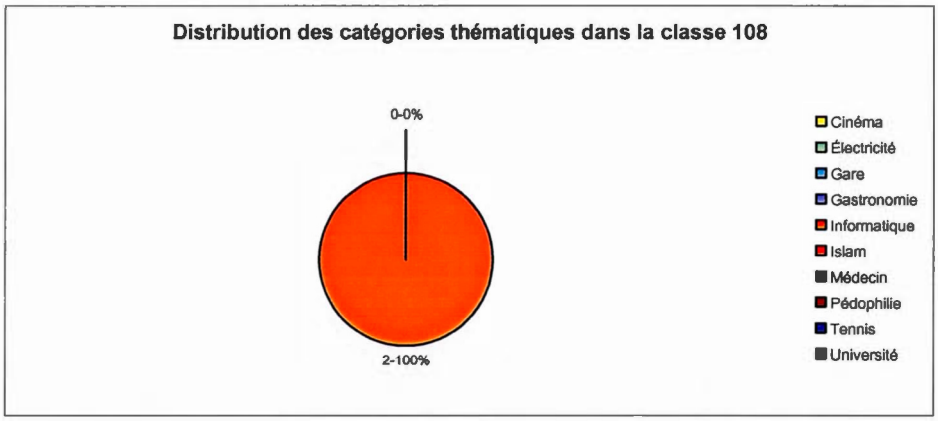

Distribution des catégories thématiques dans la classe 103
0-0%
6-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

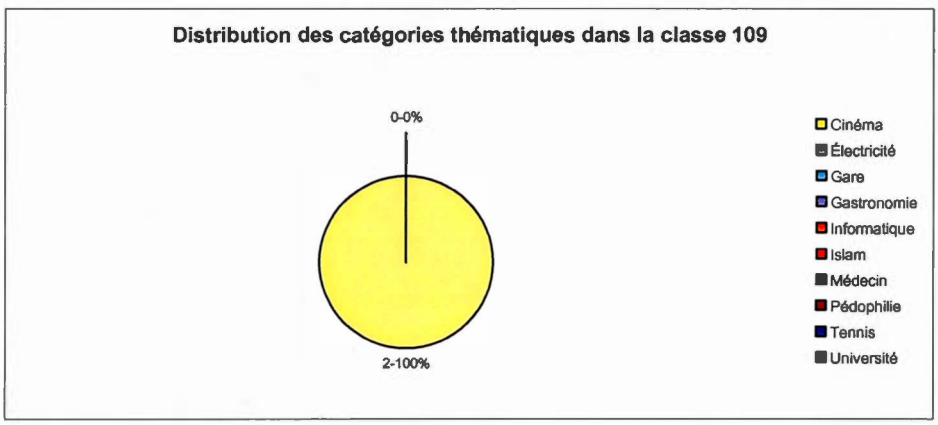

Distribution des catégories thématiques dans la classe 104
0-0%
1-14%
0-0%
6-86%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

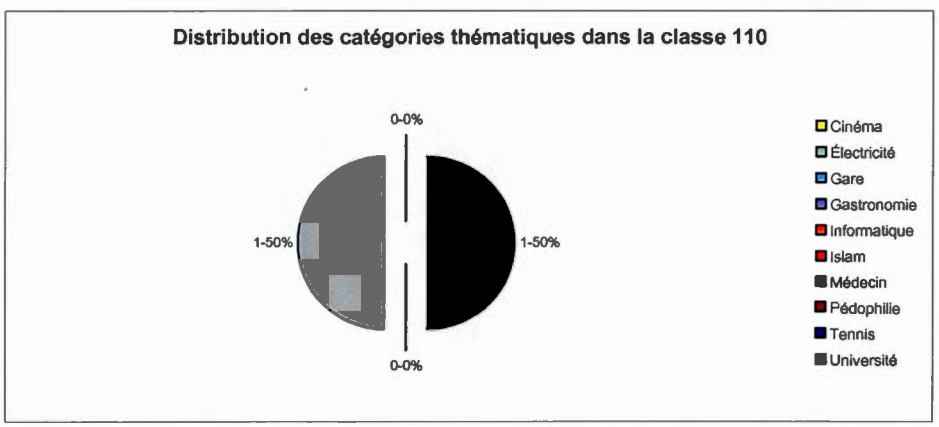

Distribution des catégories thématiques dans la classe 105
0-0%
1-25%
0-0%
3-75%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

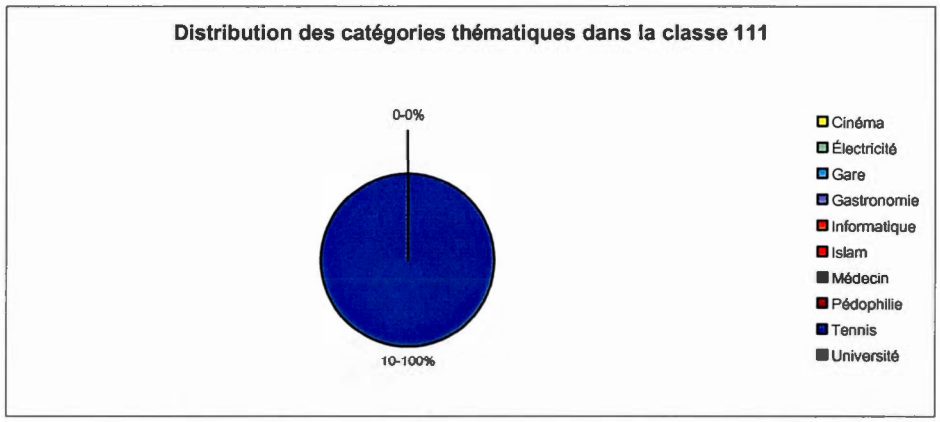

Distribution des catégories thématiques dans la classe 106
0-0%
19-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

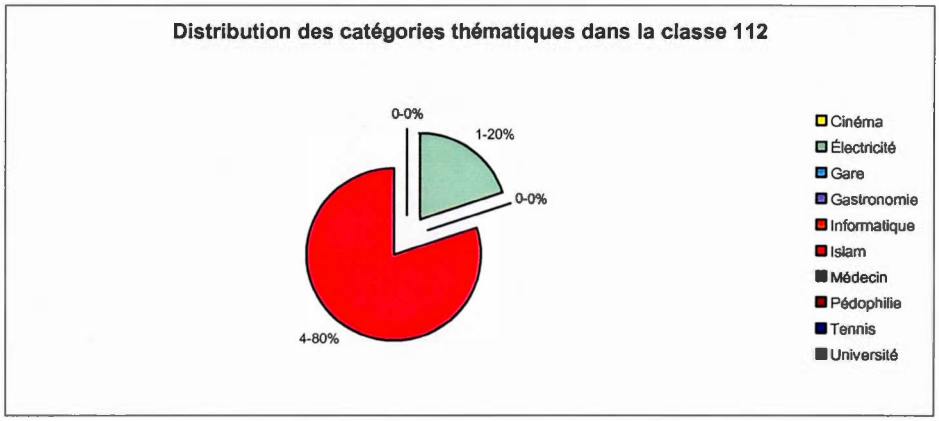

Distribution des catégories thématiques dans la classe 107
0-0%
1-33%
2-67%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

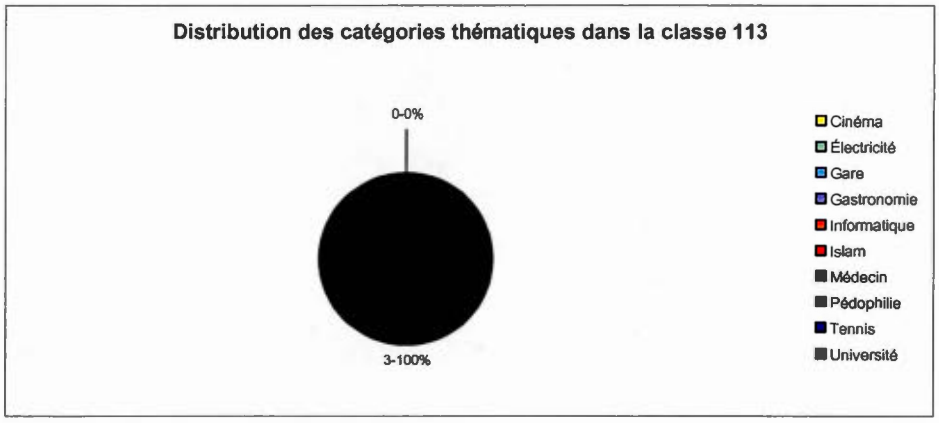

Distribution des catégories thématiques dans la classe 108
0-0%
2-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

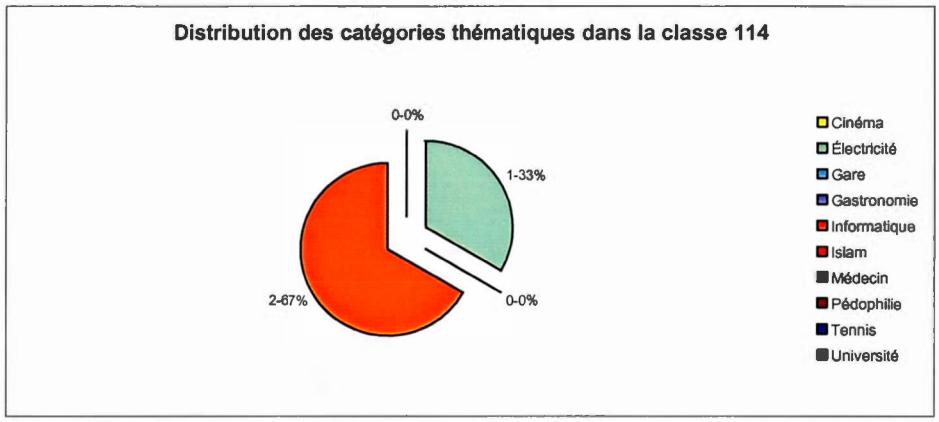

Distribution des catégories thématiques dans la classe 109
0-0%
2-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

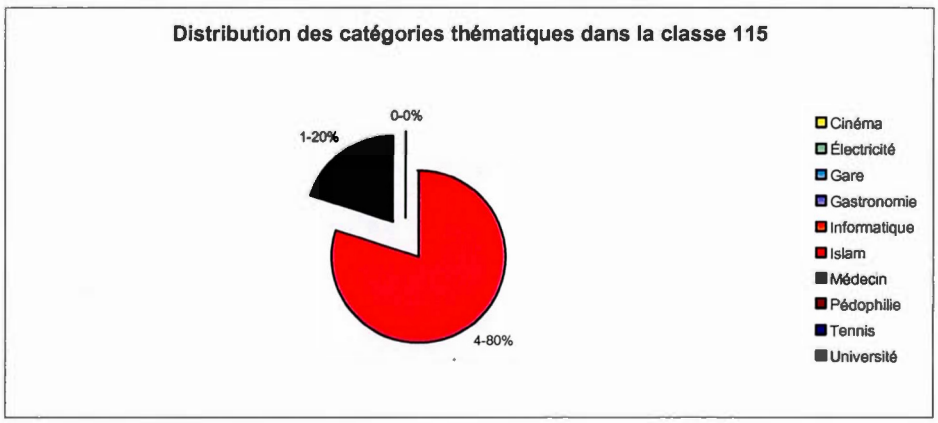

Distribution des catégories thématiques dans la classe 110
0-0%
1-50%
1-50%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

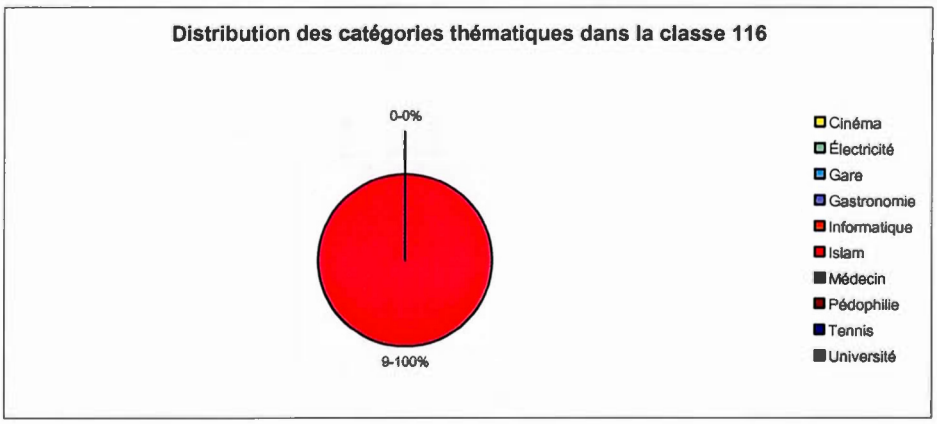

Distribution des catégories thématiques dans la classe 111
0-0%
10-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

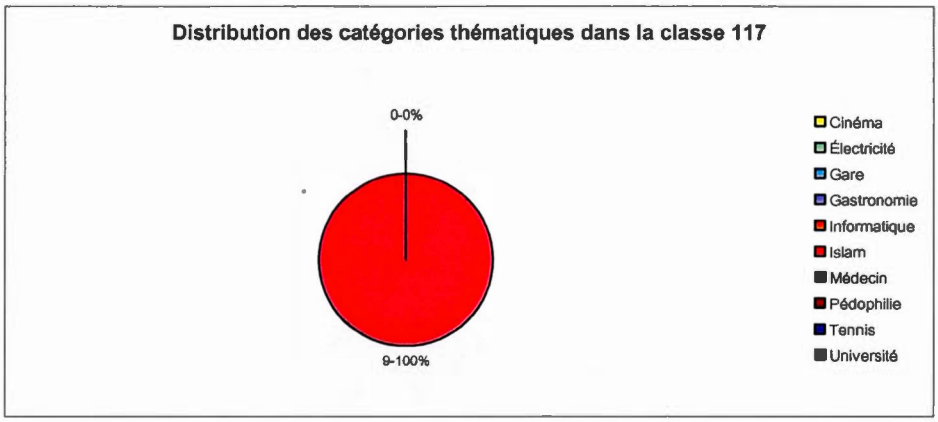

Distribution des catégories thématiques dans la classe 112
0-0%
1-20%
0-0%
4-80%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

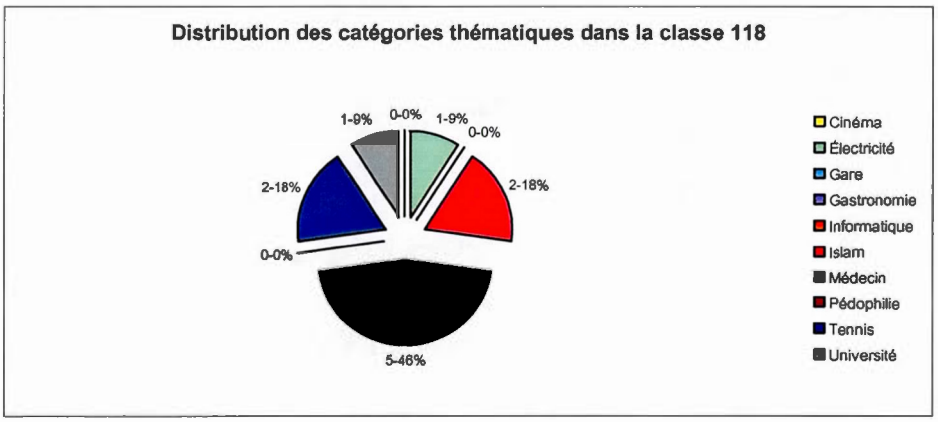

Distribution des catégories thématiques dans la classe 113
0-0%
3-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 114
0-0%
1-33%
2-67%
0-0%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 115
0-0%
1-20%
4-80%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 116
0-0%
9-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 117
0-0%
9-100%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université


Distribution des catégories thématiques dans la classe 118
1-9%
0-0%
1-9%
0-0%
2-18%
2-18%
0-0%
5-46%
Cinéma
Électricité
Gare
Gastronomie
Informatique
Islam
Médecin
Pédophilie
Tennis
Université

ANNEXE 6

ÉVALUATION DES RÉSULTATS SELON LA MESURE DE HIRST ET ST-ONGE

CLASSE	CATÉGORIE ATTRIBUÉE MANUELLEMENT	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	SCORE
001	Informatique	Système	0
		Microsoft	n/a (pas dans WORDNET)
		Ordinateur	16
002	Islam	Avocat	0
		Ministre	0
		Politique	0
003	Médecin	Médecin	16
		Généraliste	6
		Patient	2
004	Université	Professeur	4
		Festival	0
		Classe	4
005	Gastronomie	Produit	0
		Table	0
		Goût	0
006	Islam	Violence	0
		Communauté	4
		Film	0
007	Informatique	Logiciel	5
		Microsoft	n/a (pas dans WORDNET)
		Sécurité	3
008	Tennis	Foi	0
		Femme	0
		Énergie	0
009	Gare	Chemin [de fer]	0
		Police	0
		Gare	16
010	Gastronomie	Recette	0
		Cuisine	4
		Table	0
011	Médecin	Porto	0
		Avocat	3
		Cuisine	0
012	Pédophilie	Dutroux	n/a (pas dans WORDNET)

CLASSE	CATÉGORIE ATTRIBUÉE MANUELLEMENT	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	SCORE
		Dossier	0
		Nihoul	n/a (pas dans WORDNET)
013	Université	Création	3
		Société	4
		Réseau	2
014	Électricité	Réseau	0
		Électricité	16
		Électrabel	n/a (pas dans WORDNET)
015	Université	Étude	2
		Énergie	0
		Secondaire	3
016	Tennis	Tournoi	0
		Clijsters	n/a (pas dans WORDNET)
		Henin	n/a (pas dans WORDNET)
017	Pédophilie	Plainte	0
		Dossier	0
		Procureur	0
018	Islam	Attentat	0
		Musulman	16
		États-Unis	0
019	Électricité	Électrique	16
		Vélo	0
		Transport	0
020	Université	Étudier	2
		Enseignement	0
		Université	16
021	Tennis	Victoire	6
		Clijsters	n/a (pas dans WORDNET)
		Débat	0
022	Islam	Politique	0
		Pouvoir	2
		Communauté	4
023	Gastronomie	Coût	0
		Projet	2
		Prix	0
024	Gastronomie	Maison	0
		Chef	0
		Produit	0
025	Informatique	Informatique	16
		Société	3
		Projet	0
026	Électricité	Commission	0
		Loi	0
		Enseignement	0
027	Tennis	Henin	n/a (pas dans WORDNET)
		Clijsters	n/a (pas dans WORDNET)

CLASSE	CATÉGORIE ATTRIBUÉE MANUELLEMENT	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	SCORE
		Williams	n/a (pas dans WORDNET)
028	Tennis	Victoire	6
		Wiliams	n/a (pas dans WORDNET)
		Tennis	16
029	Médecin	Médical	0
		UCL	n/a (pas dans WORDNET)
		Santé	0
030	Électricité	Électrabel	n/a (pas dans WORDNET)
		Électricité	16
		Producteur	0
031	Cinéma	Producteur	0
		Film	0
		Jeu	0
032	Islam	Gouvernement	2
		Ministre	0
		Islamiste	16
033	Tennis	Clijsters	n/a (pas dans WORDNET)
		Balle	0
		Jeu	5
034	Islam	Musulman	16
		Communauté	4
		Ministre	0
035	Cinéma	Cinéma	16
		Industrie	0
		Studio	0
036	Pédophilie	Directeur	0
		Projet	0
		Enfant	0
037	Gare	Gare	16
		Quartier	0
		Voie	0
038	Pédophilie	Professionnel	0
		Avocat	0
		Dossier	0
039	Pédophilie	Procureur	0
		Avocat	0
		Bourlet	n/a (pas dans WORDNET)
040	Pédophilie	Tribunal	0
		Avocat	0
		Juge	0
041	Médecin	Médecin	16
		Médical	0
		Patient	2
042	Médecin	Santé	0
		Soin	2
		Spécialiste	4

CLASSE	CATÉGORIE ATTRIBUÉE MANUELLEMENT	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	SCORE
043	Islam	Musulman	16
		Mosquée	0
		Islamiste	16
044	Université	Enseignement	0
		Université	16
		Recherche	0
045	Gare	Train	0
		Vie	0
		Projet	0
046	Islam	Arabe	0
		Monde	2
		Valeur	0
047	Cinéma	Acteur	0
		Monde	0
		Confiance	0
048	Tennis	Sport	5
		Tennis	16
		Stade	0
049	Électricité	Consommateur	0
		Compte	0
		Monde	2
050	Université	Laboratoire	0
		ULB	n/a (pas dans WORDNET)
		Recherche	0
051	Gare	SNCB	n/a (pas dans WORDNET)
		Passage	2
		Collège	0
052	Médecin	Santé	0
		Politique	0
		Soin	2
053	Islam	Musulman	16
		Politique	0
		Monde	2
054	Islam	Religieux	4
		AKP	n/a (pas dans WORDNET)
		Média	0
055	Tennis	Set	3
		Match	3
		Jeu	5
056	Pédophilie	Collège	0
		Procès	0
		Enfant	0
057	Université	Université	16
		Débat	0
		Formation	0
058	Informatique	Internet	5

CLASSE	CATÉGORIE ATTRIBUÉE MANUELLEMENT	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	SCORE
		Informatique	16
		Page	2
059	Islam	Justice	2
		International	0
		Équipe	2
060	Cinéma	Film	0
		Scène	0
		Image	0
061	Informatique	Logiciel	5
		Réseau	2
		Virus	2
062	Médecin	Professeur	0
		Recherche	0
		Islamiste	2
063	Cinéma	Film	0
		Scène	0
		Numérique	0
064	Université	Vélo,	0
		Sécurité	0
		Qualité	0
	Électricité	Vélo,	0
		Sécurité	2
		Qualité	0
065	Islam	Police	2
		Quartier	0
		Fédéral	0
066	Pédophilie	Loi	0
		Enseignement	0
		Enfant	0
	Médecin	Loi	0
		Enseignement	3
		Enfant	3
067	Pédophilie	Victime	0
		Autorité	0
		Dossier	0
068	Informatique	Technologie	0
		Entreprise	3
		Secteur	2
	Électricité	Technologie	0
		Entreprise	0
		Secteur	0
069	Gare	Gare	16
		Train	0
		SNCB	n/a (pas dans WORDNET)
070	Tennis	Masters	0
		Gagner	4

CLASSE	CATÉGORIE ATTRIBUÉE MANUELLEMENT	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	SCORE
		Mondial	0
071	Islam	Mosquée	0
		Quartier	0
		Turc	0
072	Islam	Musulman	16
		Islam	16
		Communauté	4
073	Cinéma	Film	0
		Cinéma	16
		Scène	0
074	Tennis	Match	3
		Jeu	5
		Joueur	0
075	Tennis	Henin	n/a (pas dans WORDNET)
		Match	3
		Coup	0
076	Islam	Islamiste	16
		Ville	0
		Coup	0
077	Islam	Islamique	16
		Art	0
		Projet	0
078	Université	Formation	0
		Organisation	5
		Université	16
079	Tennis	Victoire	6
		Tournoi	0
		Coup	0
080	Gastronomie	Cuisine	4
		Restaurer	0
		Chef	0
081	Pédophilie	Juge	0
		Avocat	0
		Client	0
082	Médecin	Université	0
		Étude	0
		Association	0
	Université	Université	16
		Étude	2
		Association	0
083	Électricité	Nucléaire	0
		Réaction	4
		Produit	2
084	Informatique	Informatique	16
		Réseau	2
		Technologie	0

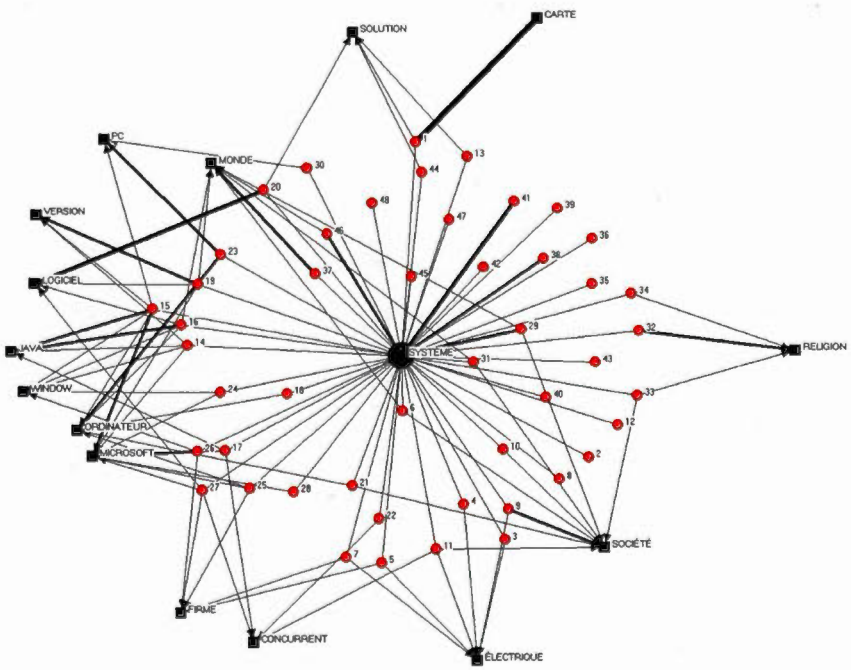
CLASSE	CATÉGORIE ATTRIBUÉE MANUELLEMENT	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	SCORE
085	Médecin	Prison	0
		Vie	3
		Malade	2
086	Islam	Gouvernement	2
		Politique	0
		Chef	3
087	Informatique	Scientifique	0
		Machine	4
		Université	0
	Médecin	Scientifique	0
		Machine	3
		Université	0
	Pédophilie	Scientifique	0
		Machine	0
		Université	0
	Université	Scientifique	0
		Machine	3
		Université	16
088	Médecin	Malade	2
		Vie	3
		Patient	2
089	Islam	Arabe	0
		Politique	0
		Monde	2
090	Tennis	Vainqueur	0
		Femme	0
		Vie	2
091	Islam	Religion	5
		Monde	2
		Islam	16
092	Tennis	Tennis	16
		Saison	2
		Coupe	0
093	Tennis	Terrain	0
		Terre	0
		Territoire	0
094	Médecin	Docteur	16
		Cas	3
		Bourgmestre	n/a (pas dans WORDNET)
095	Gastronomie	Pomme	0
		Fruit	0
		Produit	0
096	Pédophilie	Dénoncer	0
		Chef	0
		Judiciaire	0
097	Informatique	Travailler	0

CLASSE	CATÉGORIE ATTRIBUÉE MANUELLEMENT	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	SCORE
		Ordinateur	16
		Texte	0
098	Islam	Politique	0
		Islamique	16
		Religieux	4
099	Tennis	Sport	5
		Idée	0
		Résultat	3
100	Médecin	Médecin	16
		Étude	0
		Santé	0
101	Islam	Droit	2
		Gouvernement	2
		Ministre	0
102	Informatique	Ordinateur	16
		Informatique	16
		Sécurité	3
103	Gastronomie	Menu	0
		Carte	0
		Table	0
104	Université	Étudier	2
		Professeur	4
		Programme	0
105	Université	Foi	4
		Dossier	0
		Recherche	0
106	Gare	Gare	16
		SNCB	n/a (pas dans WORDNET)
		Parking	0
107	Médecin	Victime	3
		Plainte	0
		Médecin	16
108	Informatique	Numérique	0
		n/a	n/a
		n/a	n/a
109	Cinéma	Scène	0
		Guerre	0
		Passage	3
	Médecin	Scène	2
		Guerre	0
		Passage	2
110	Université	Magistrat	0
		Police	2
		Justice	0
111	Tennis	Clijsters	n/a (pas dans WORDNET)
		Masters	0

CLASSE	CATÉGORIE ATTRIBUÉE MANUELLEMENT	MOTS THÉMATIQUES CANDIDATS (FRANÇAIS)	SCORE
		Henin	n/a (pas dans WORDNET)
112	Islam	Conseil	0
		Gouvernement	2
		Ministre	0
113	Médecin	Scientifique	0
		Médical	0
		Famille	2
114	Informatique	Système	0
		Informatique	16
		Client	4
115	Islam	Droit	2
		Foi	5
		Pouvoir	2
116	Islam	Musulman	16
		Religion	5
		Islam	16
117	Islam	Arabe	0
		Ligue	2
		Islamiste	16
118	Médecin	Médecine	5
		Recherche	0
		Médical	0

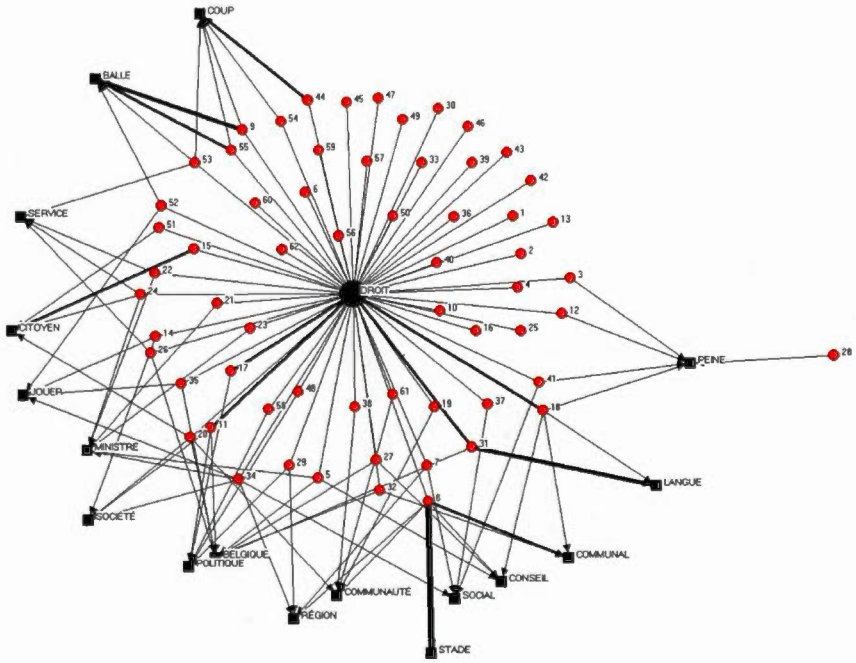
ANNEXE 7

REPRÉSENTATIONS GRAPHIQUES DU LEXIQUE DE CHAQUE CLASSE[1]

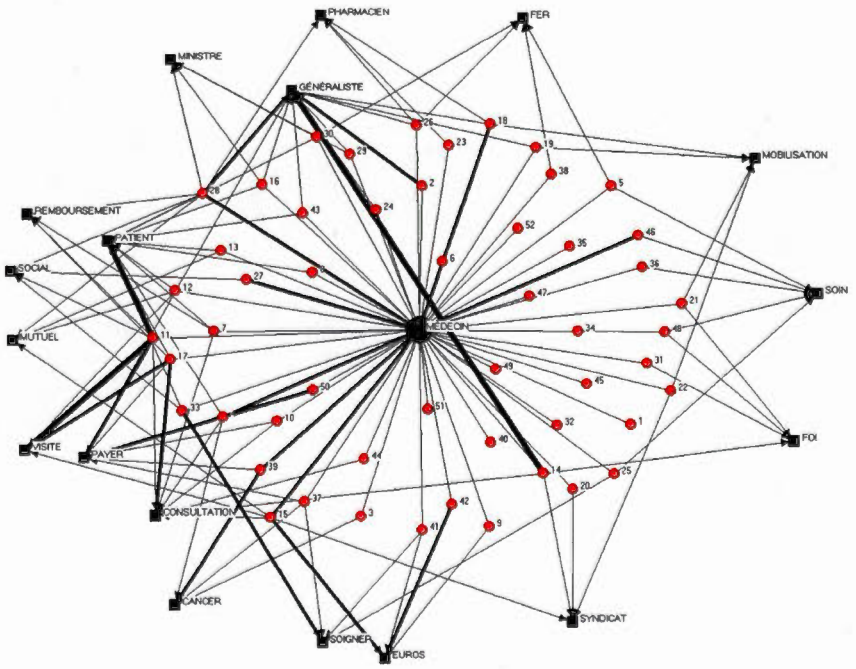


Représentation graphique du lexique de la classe 1.

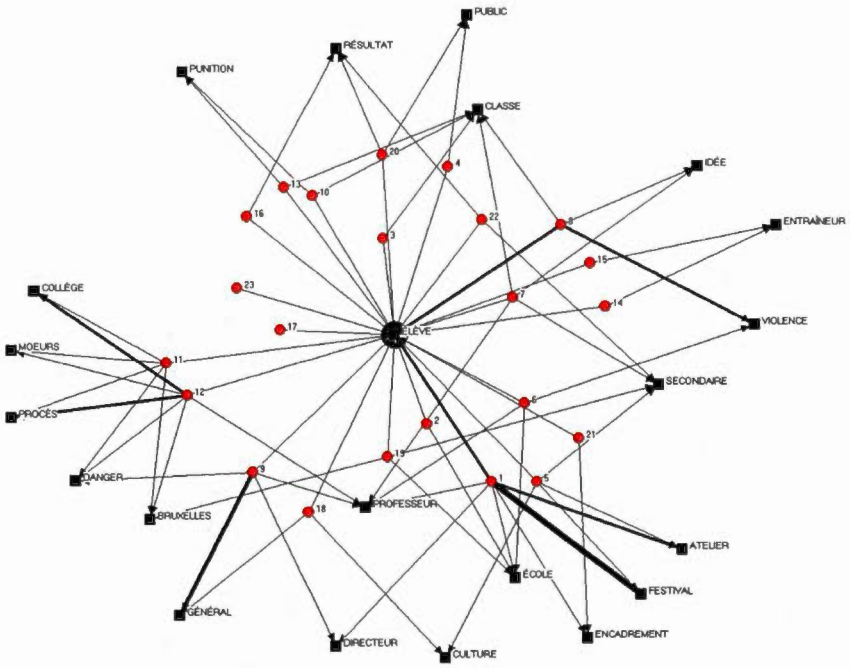
[1] Légende pour l'ensemble des figures de cette annexe : ● = Segment et ■ = Mot.



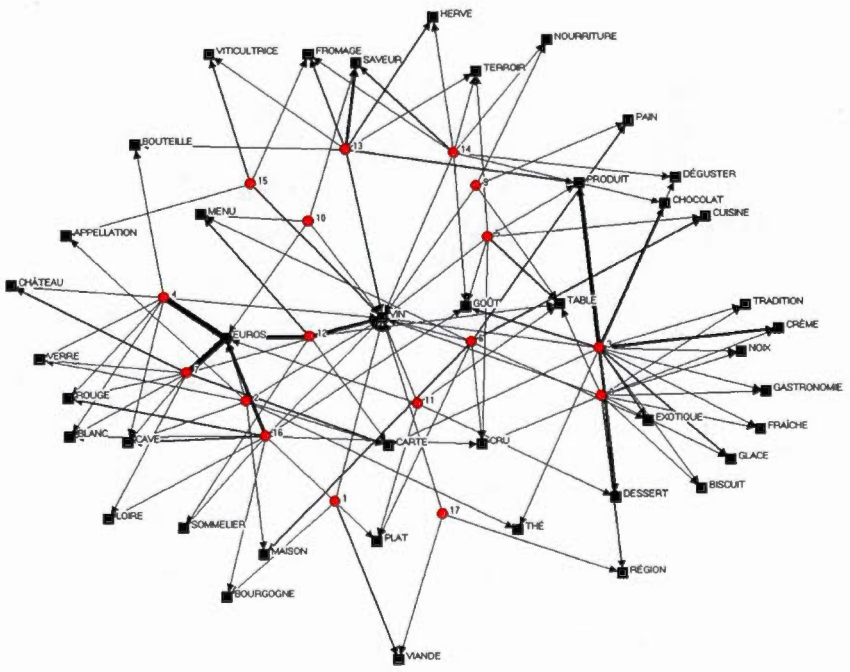
Représentation graphique du lexique de la classe 2.



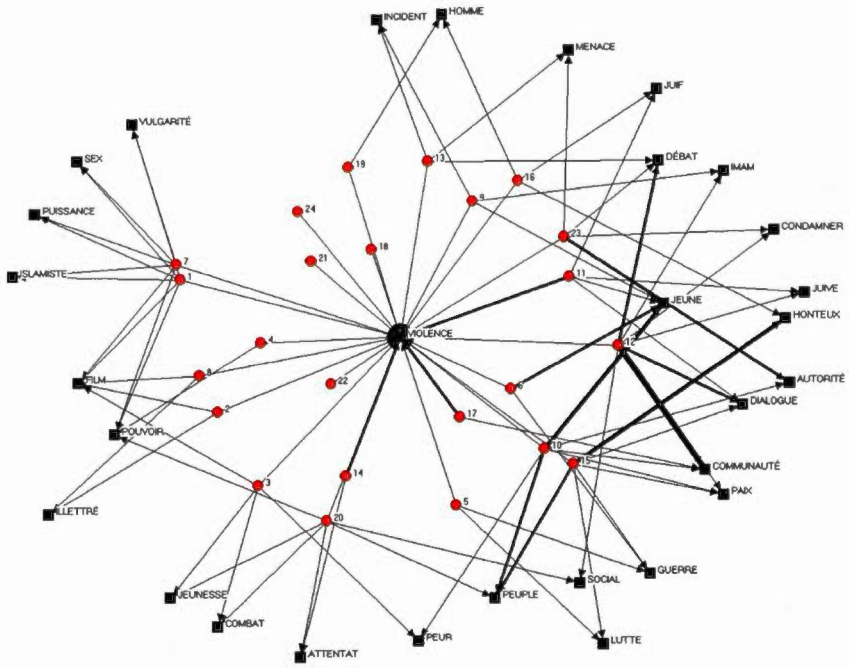
Représentation graphique du lexique de la classe 3.



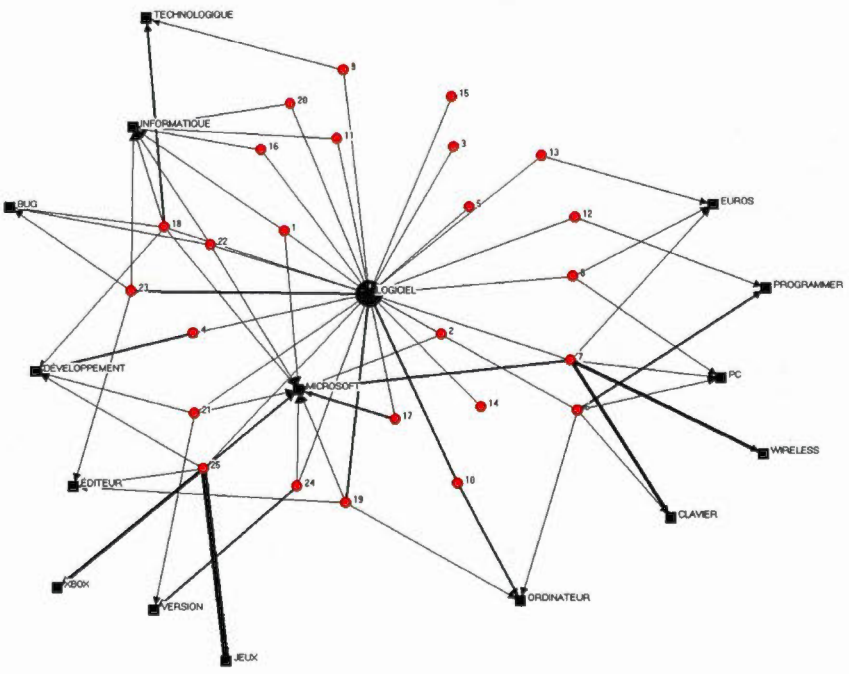
Représentation graphique du lexique de la classe 4.



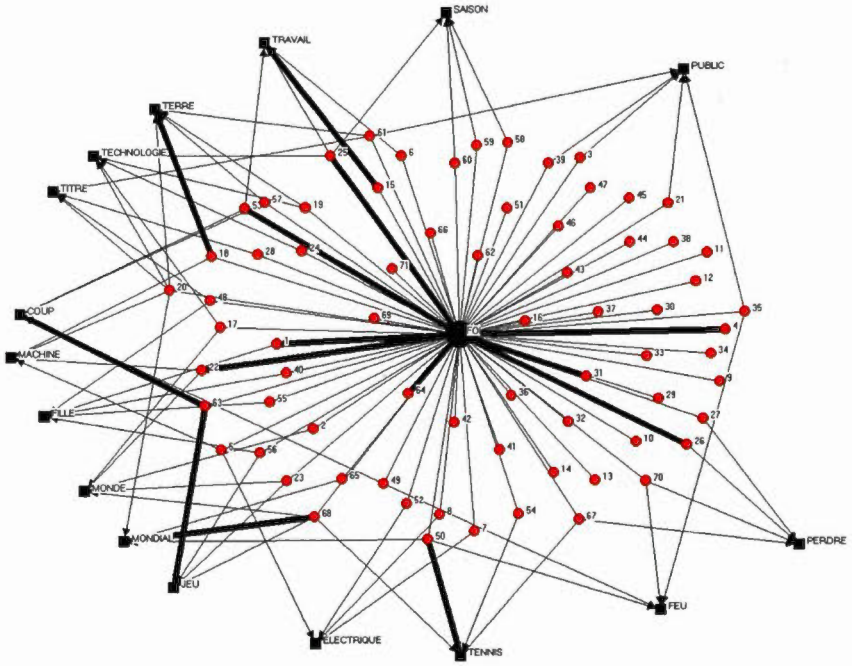
Représentation graphique du lexique de la classe 5.



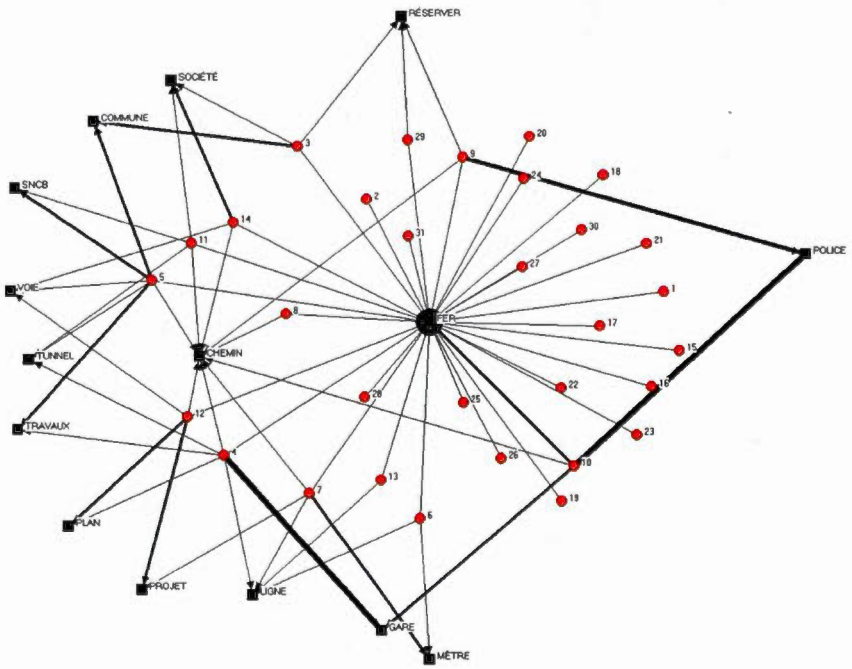
Représentation graphique du lexique de la classe 6.



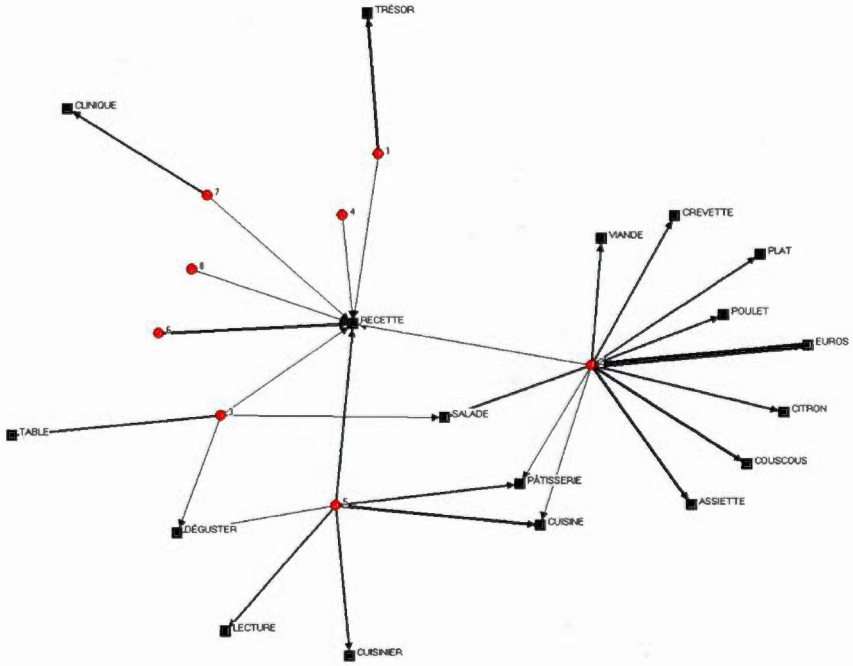
Représentation graphique du lexique de la classe 7.



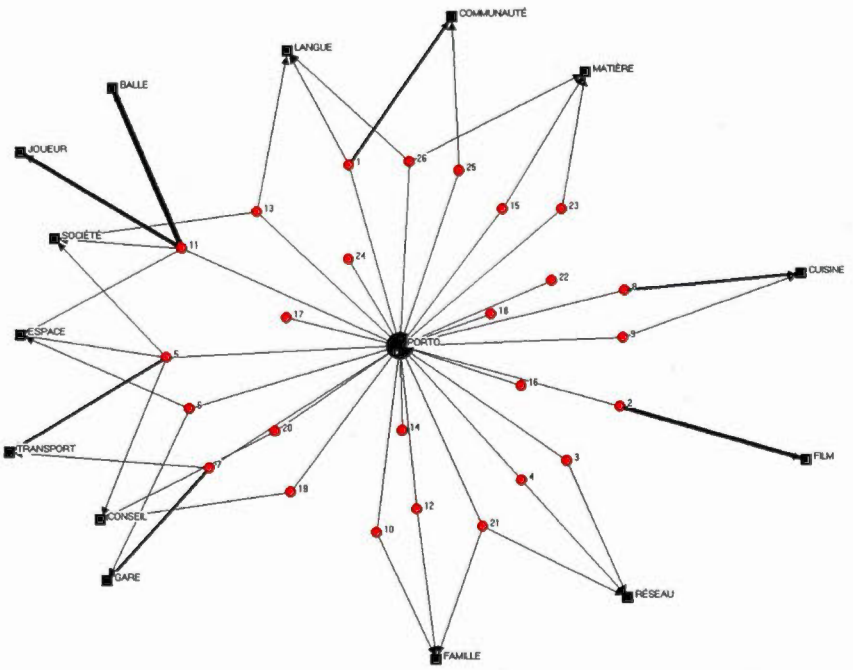
Représentation graphique du lexique de la classe 8.



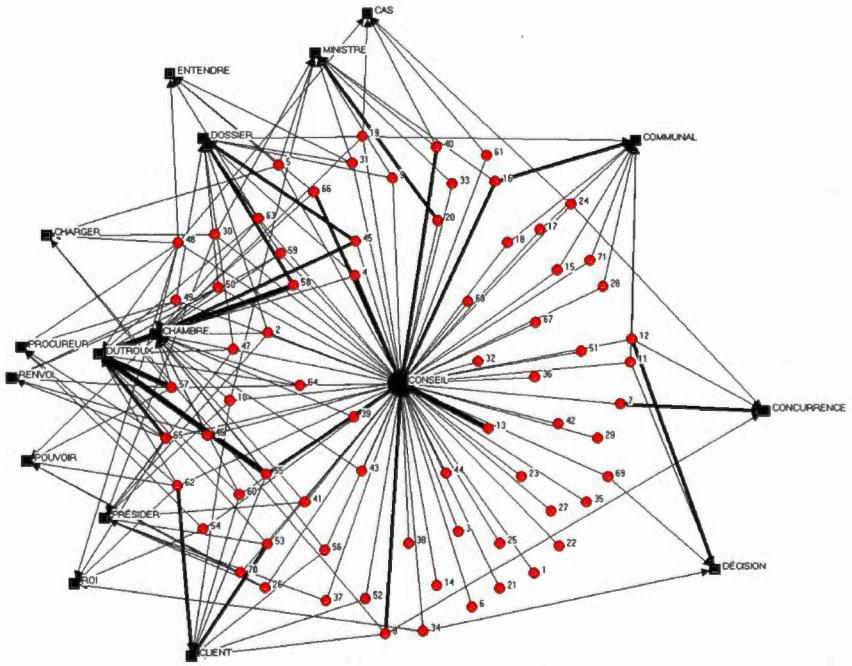
Représentation graphique du lexique de la classe 9.



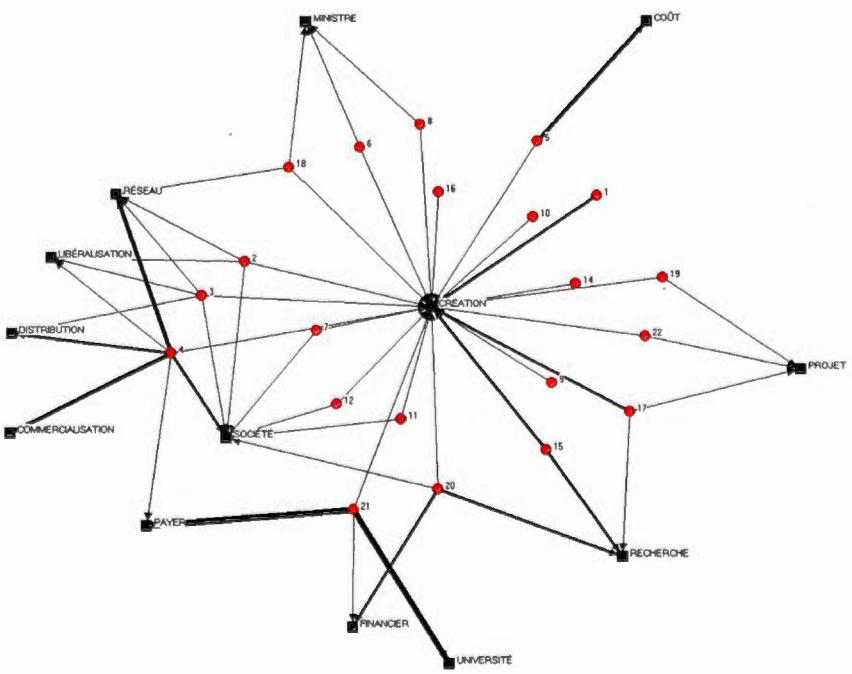
Représentation graphique du lexique de la classe 10.



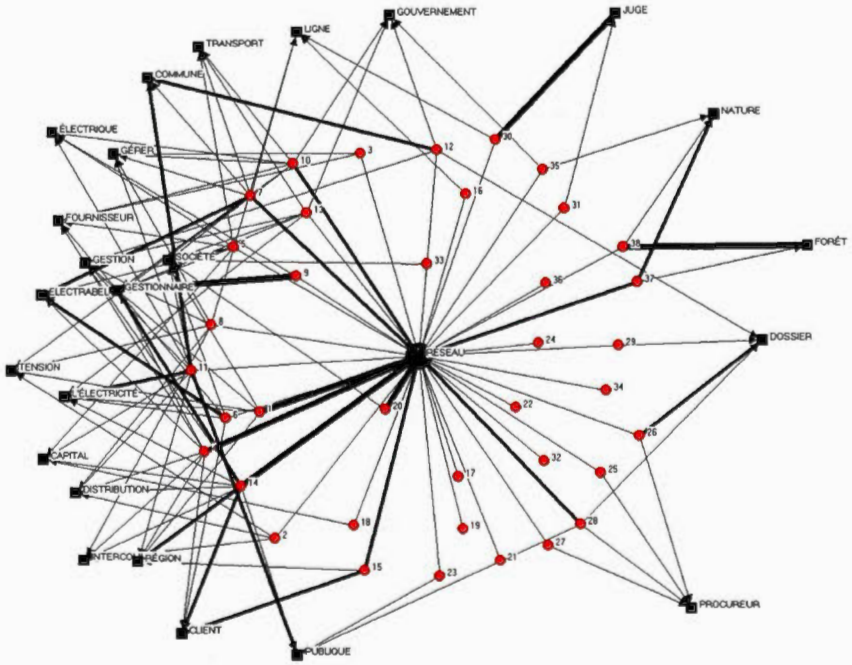
Représentation graphique du lexique de la classe 11.



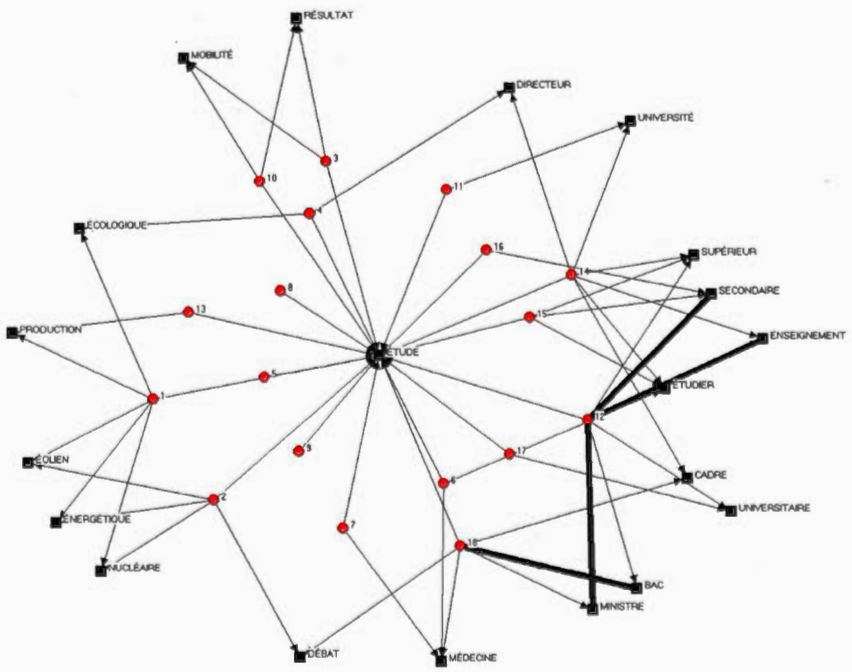
Représentation graphique du lexique de la classe 12.



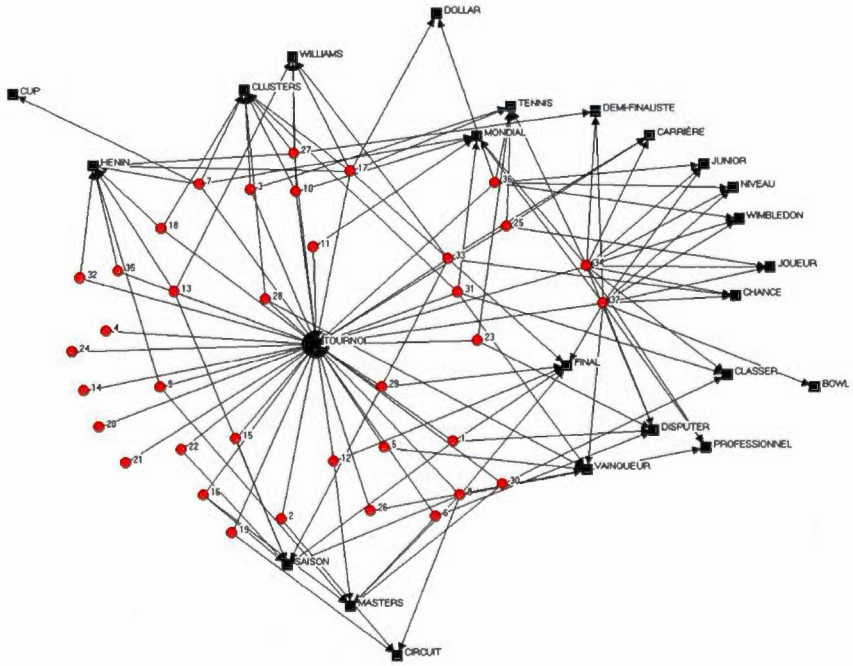
Représentation graphique du lexique de la classe 13.



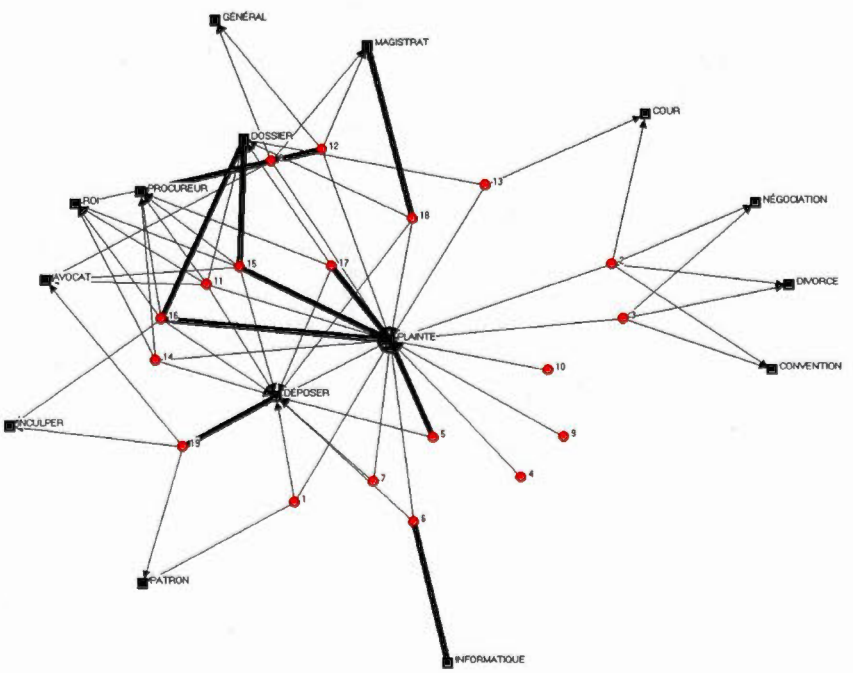
Représentation graphique du lexique de la classe 14.



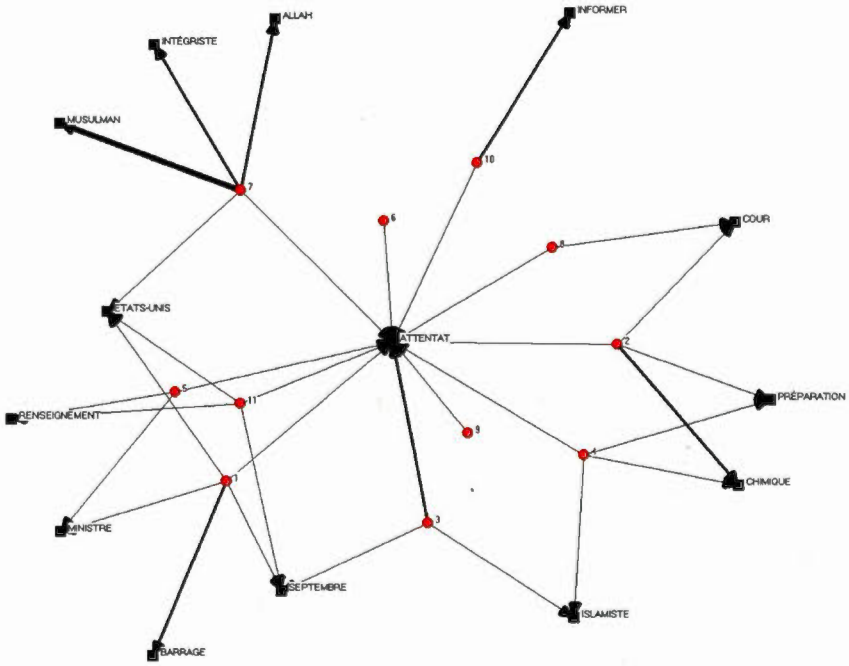
Représentation graphique du lexique de la classe 15.



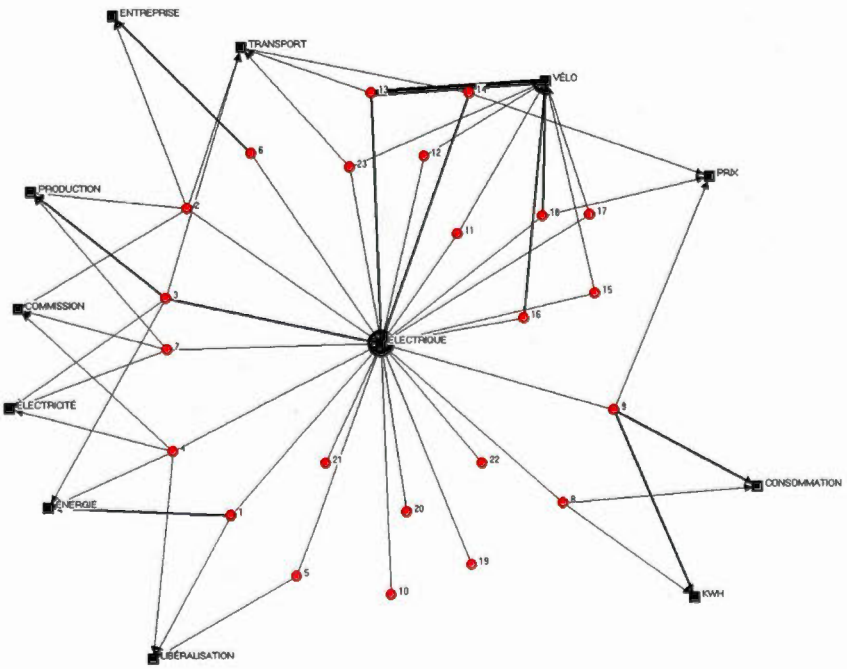
Représentation graphique du lexique de la classe 16.



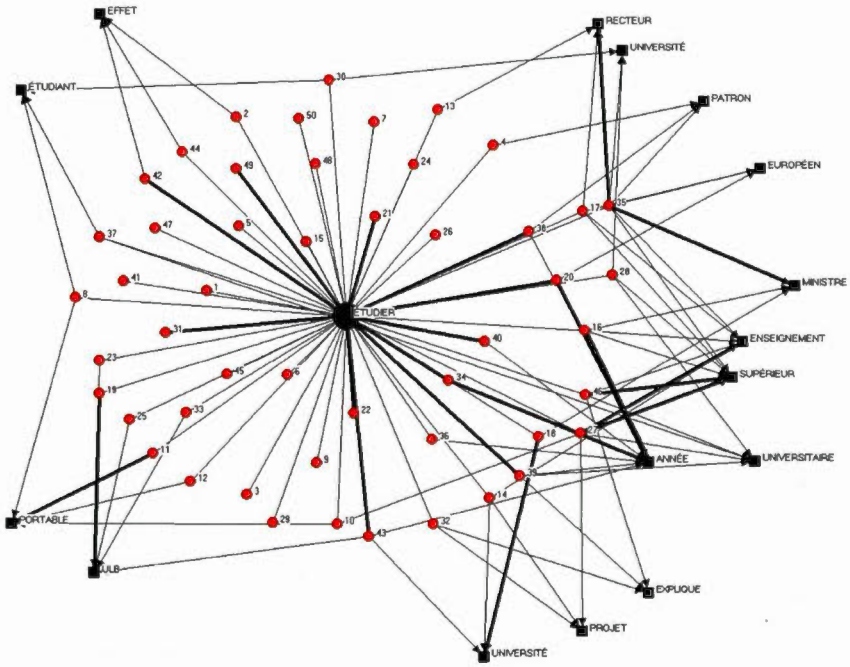
Représentation graphique du lexique de la classe 17.



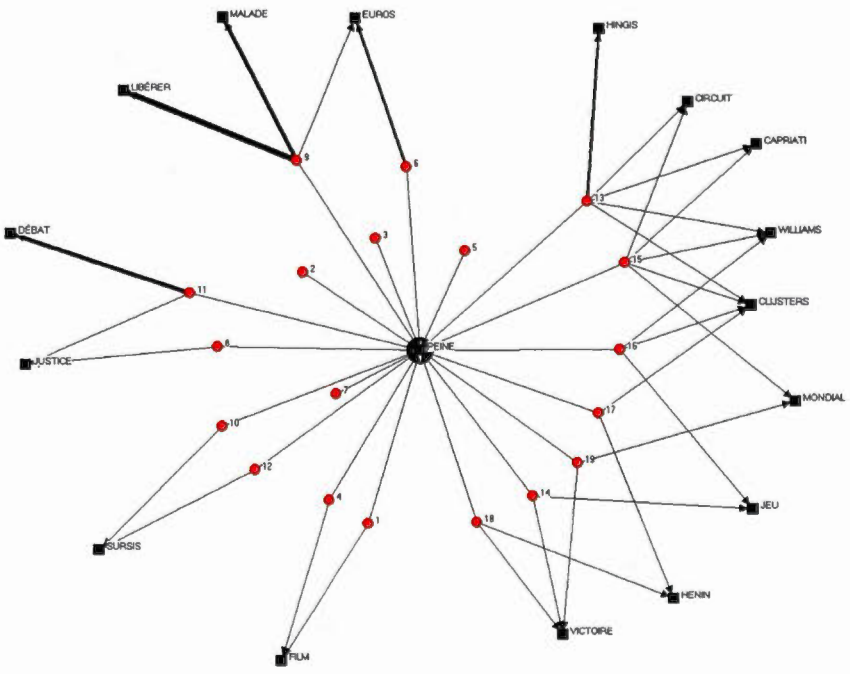
Représentation graphique du lexique de la classe 18.



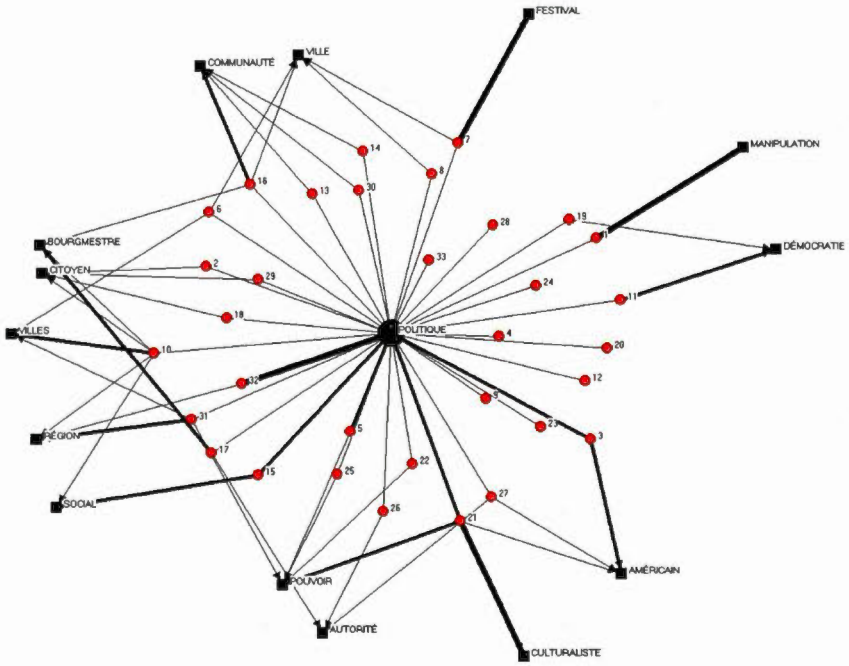
Représentation graphique du lexique de la classe 19.



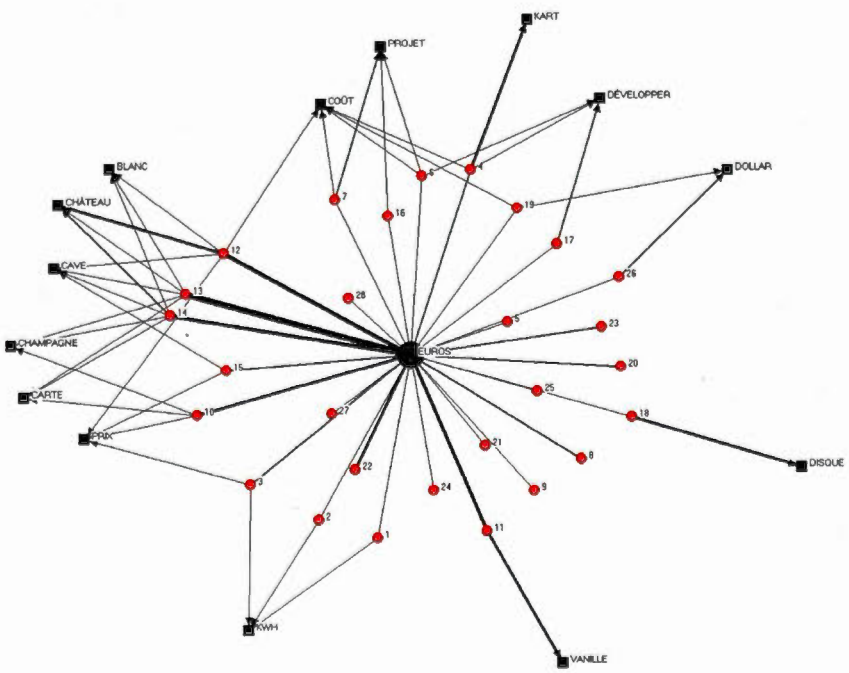
Représentation graphique du lexique de la classe 20.



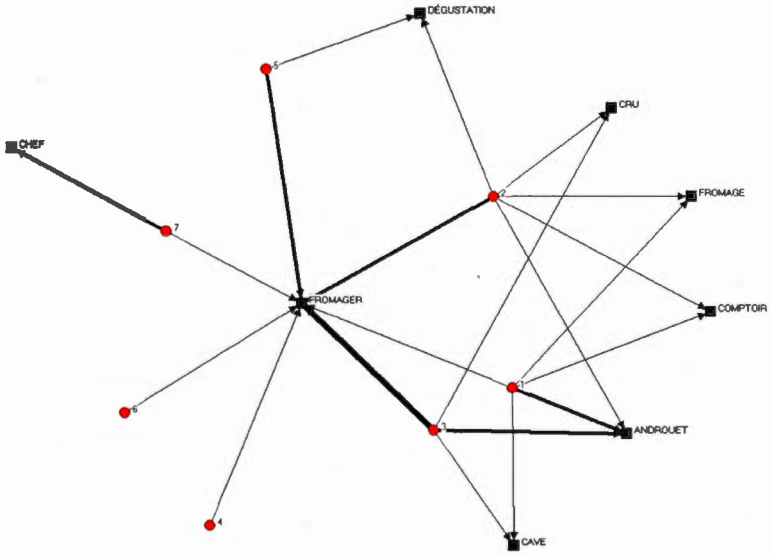
Représentation graphique du lexique de la classe 21.



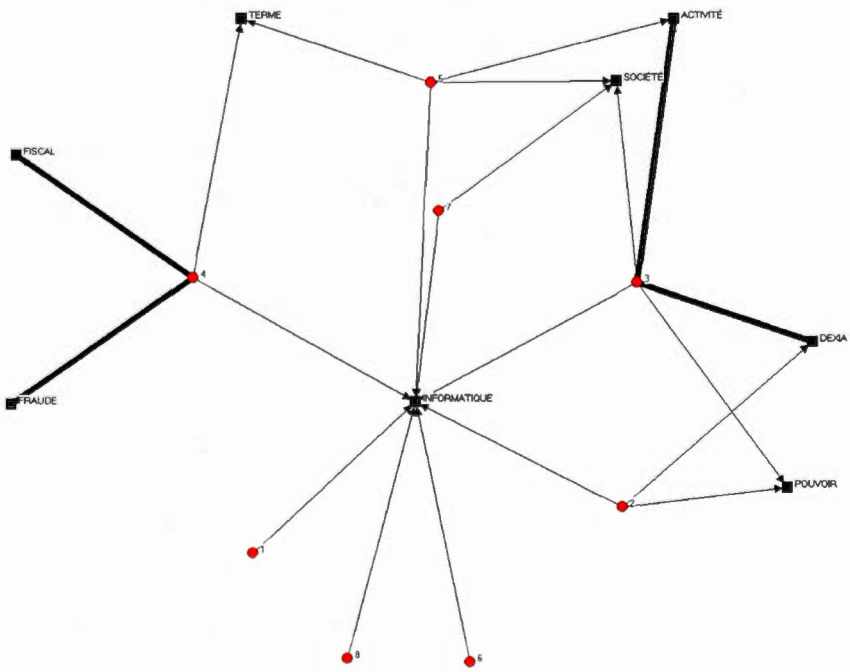
Représentation graphique du lexique de la classe 22.



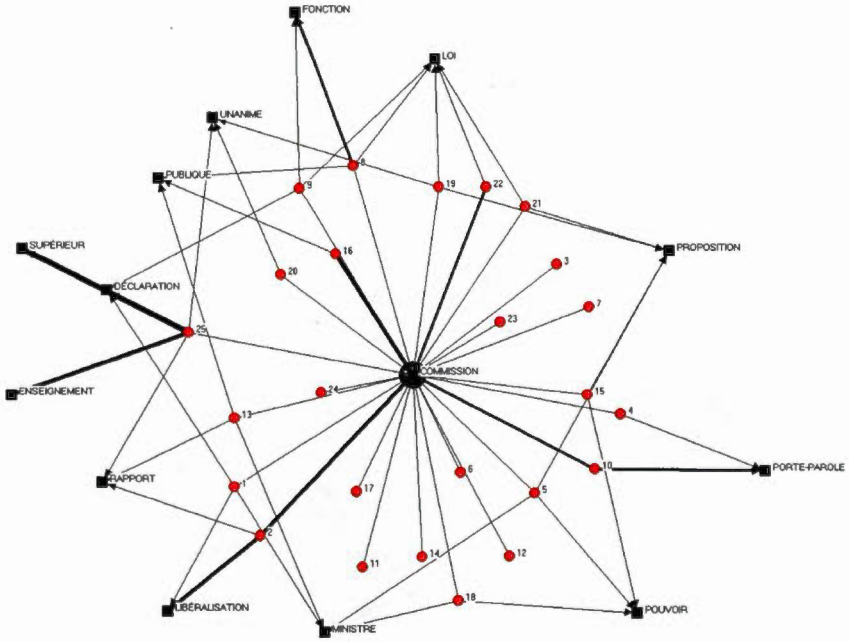
Représentation graphique du lexique de la classe 23.



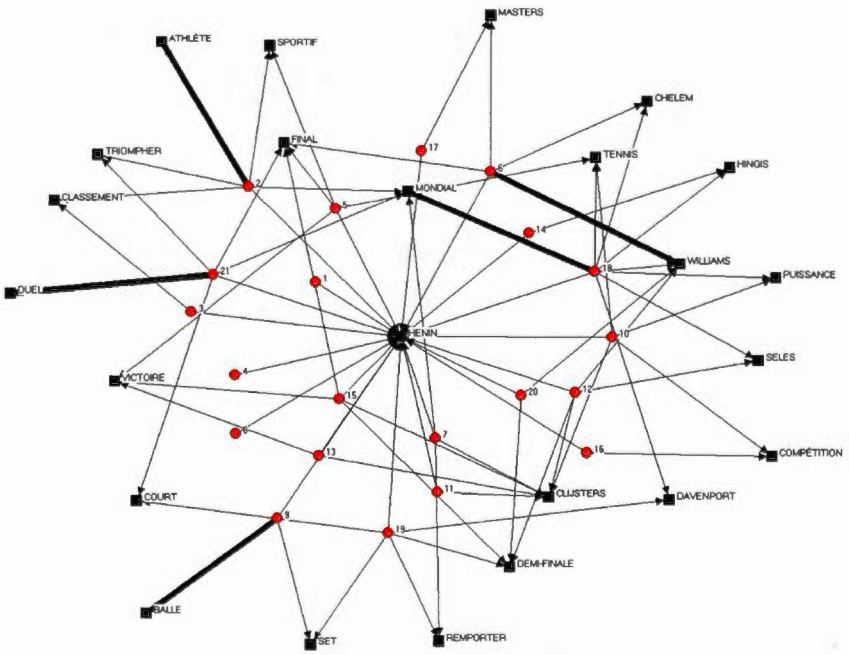
Représentation graphique du lexique de la classe 24.



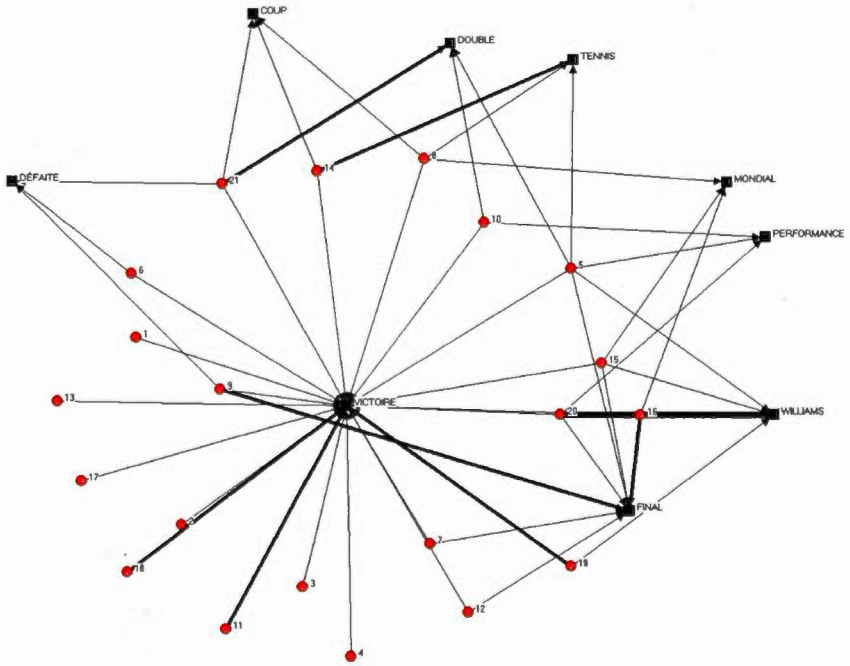
Représentation graphique du lexique de la classe 25.



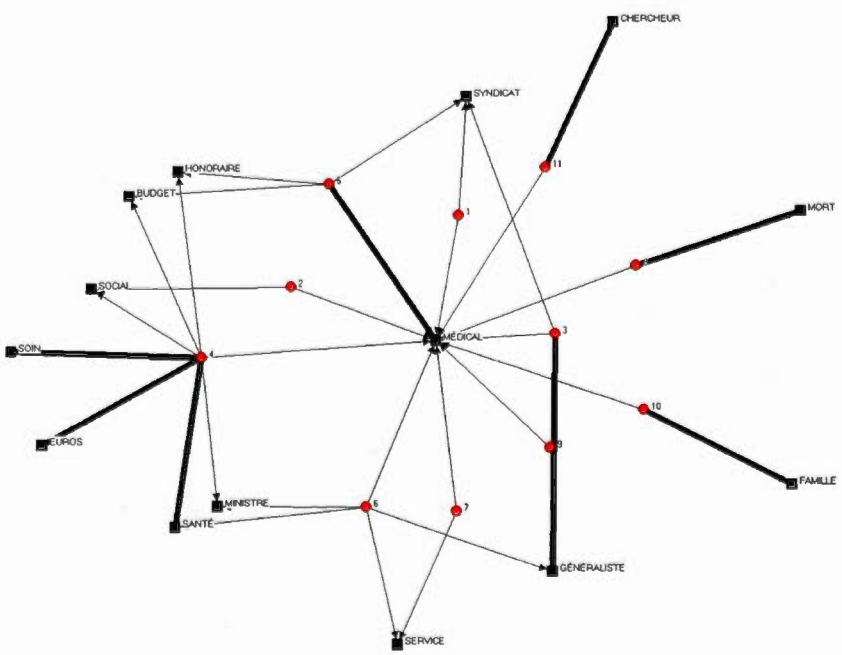
Représentation graphique du lexique de la classe 26.



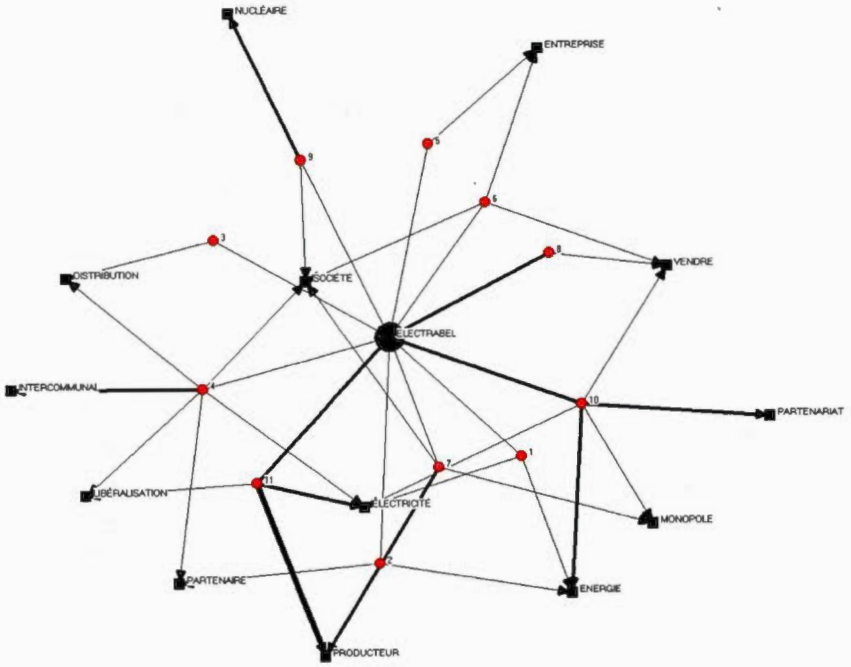
Représentation graphique du lexique de la classe 27.



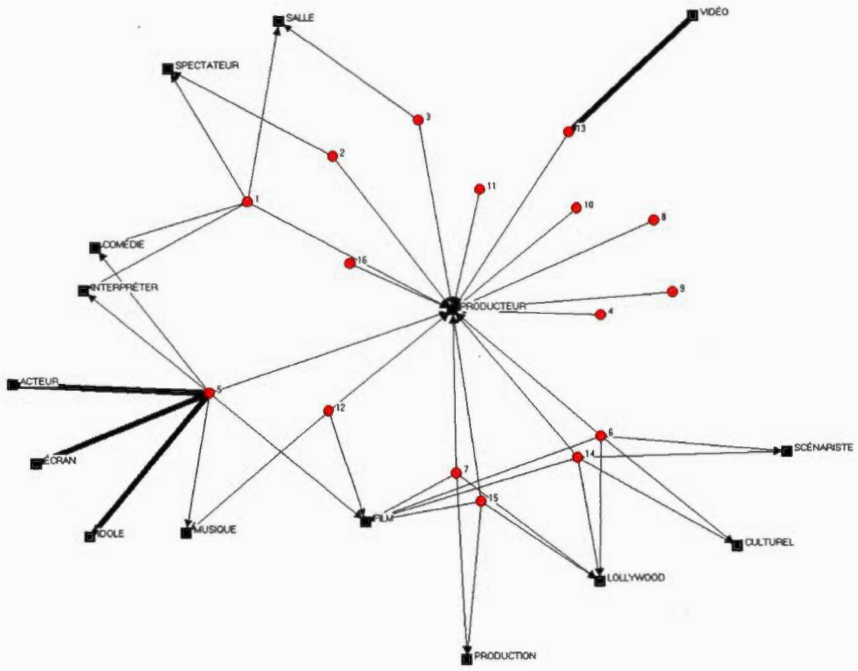
Représentation graphique du lexique de la classe 28.



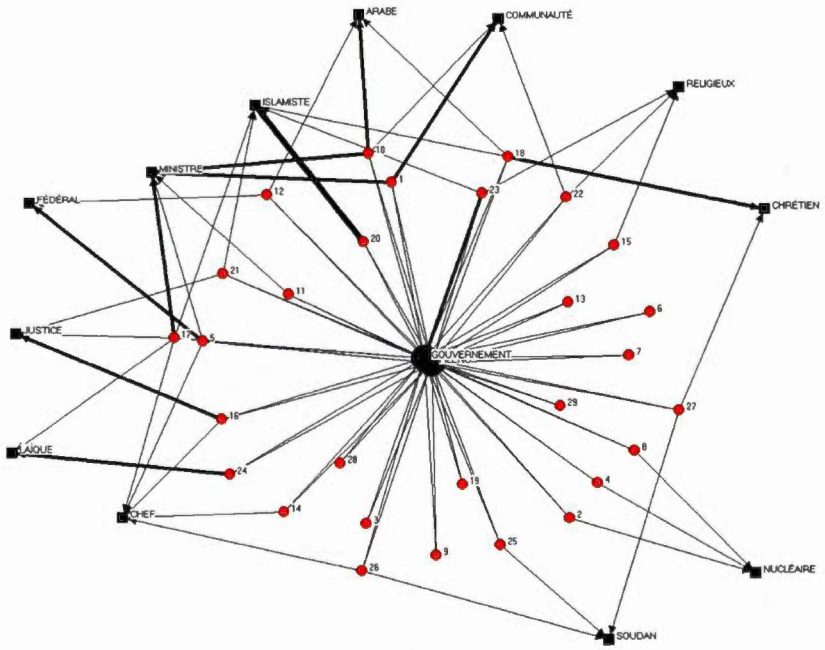
Représentation graphique du lexique de la classe 29.



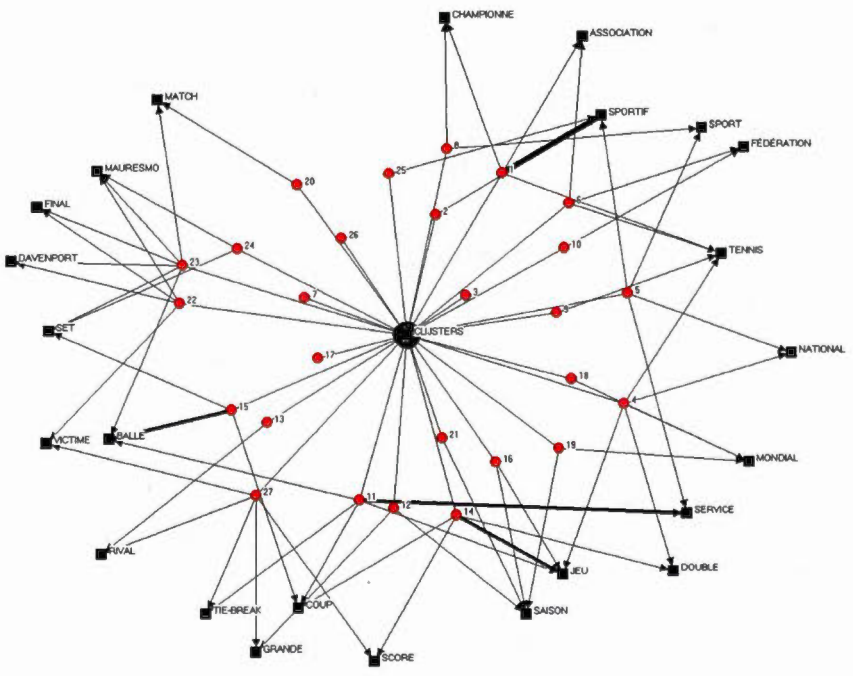
Représentation graphique du lexique de la classe 30.



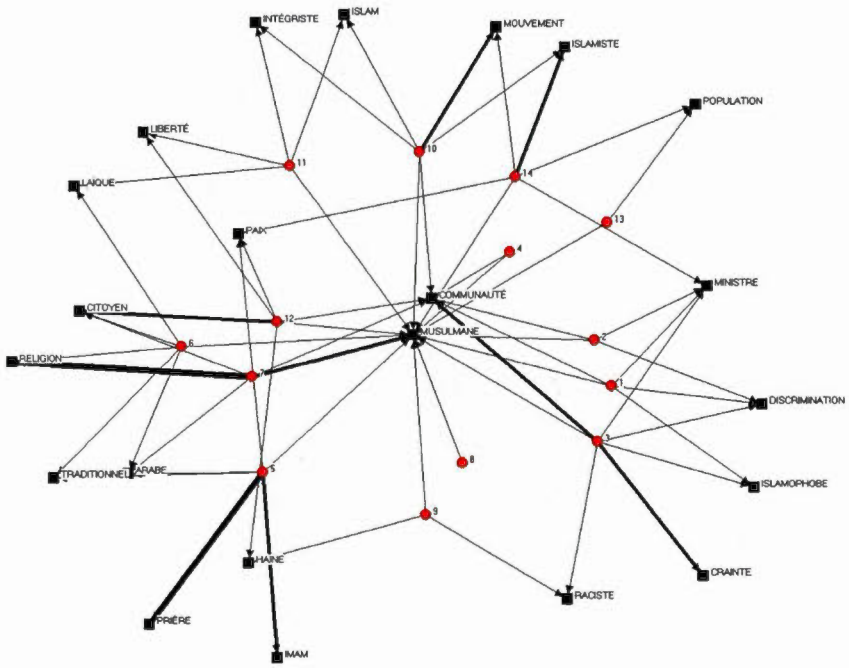
Représentation graphique du lexique de la classe 31.



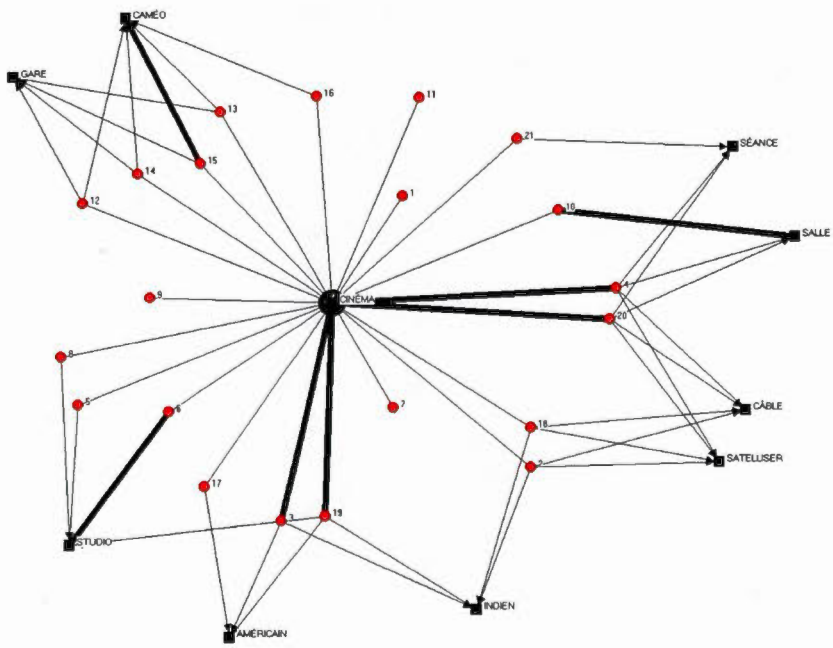
Représentation graphique du lexique de la classe 32.



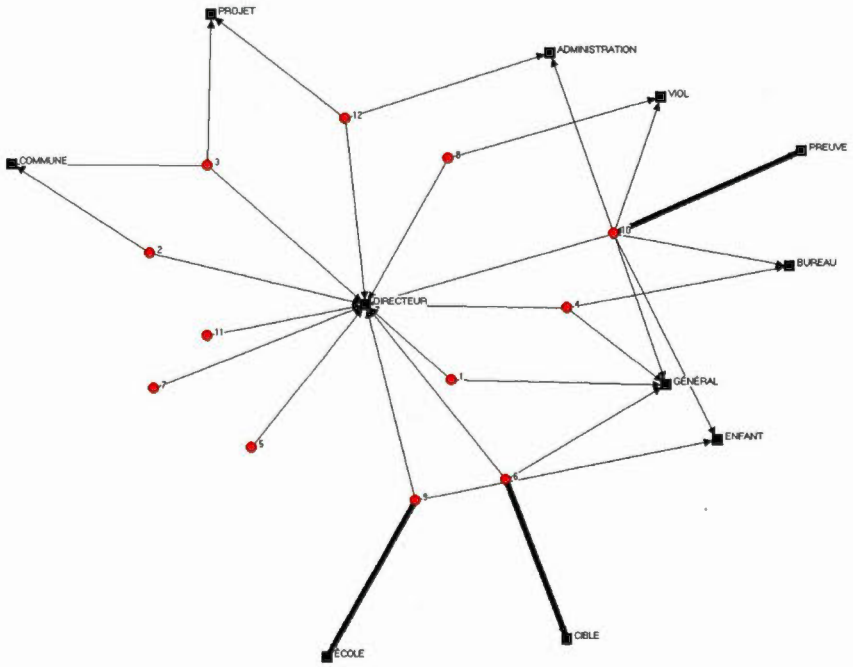
Représentation graphique du lexique de la classe 33.



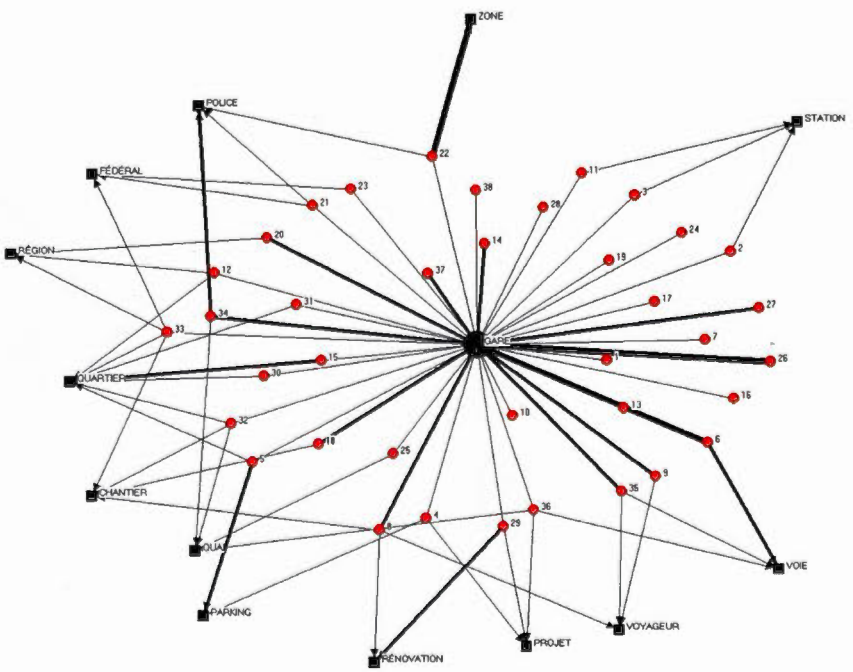
Représentation graphique du lexique de la classe 34.



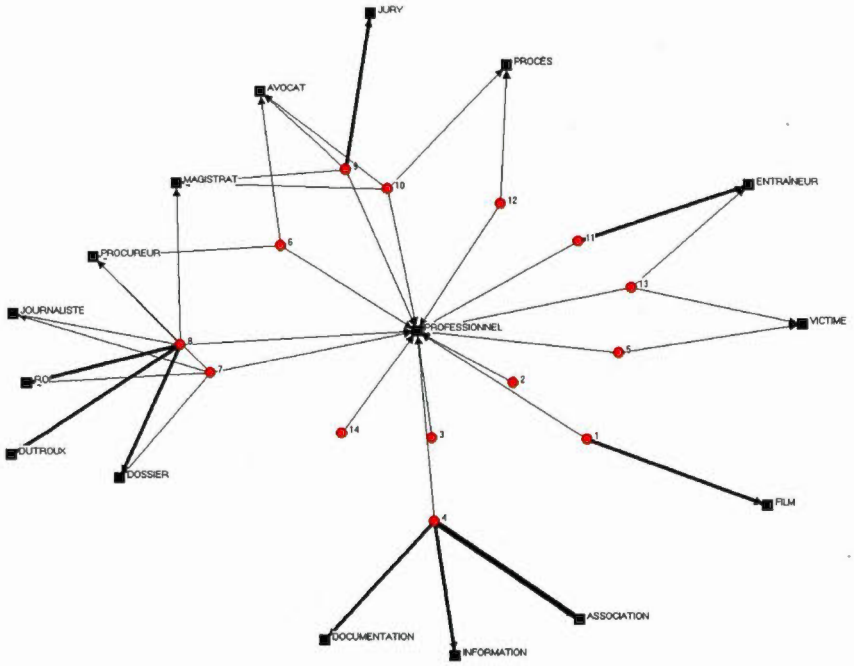
Représentation graphique du lexique de la classe 35.



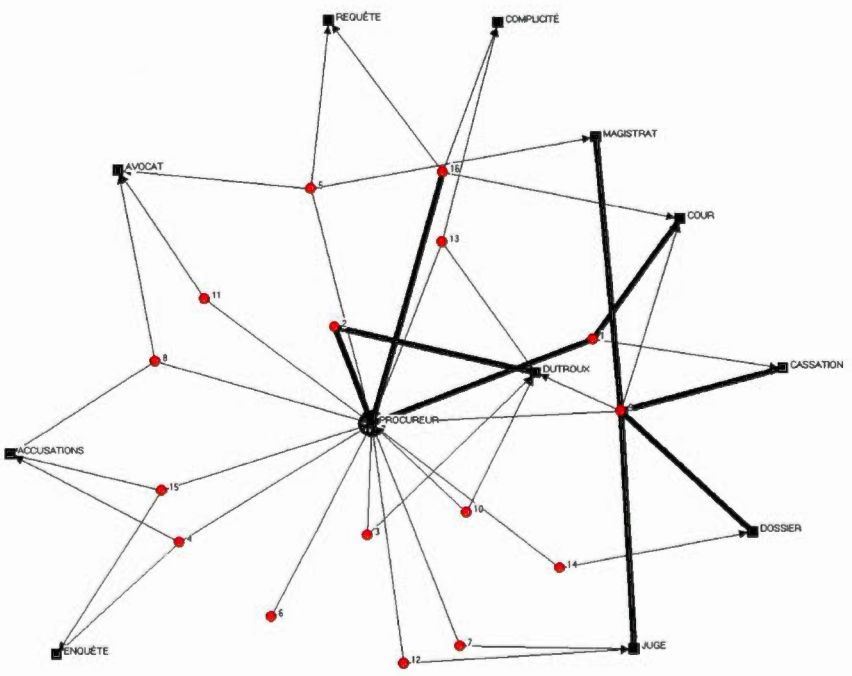
Représentation graphique du lexique de la classe 36.



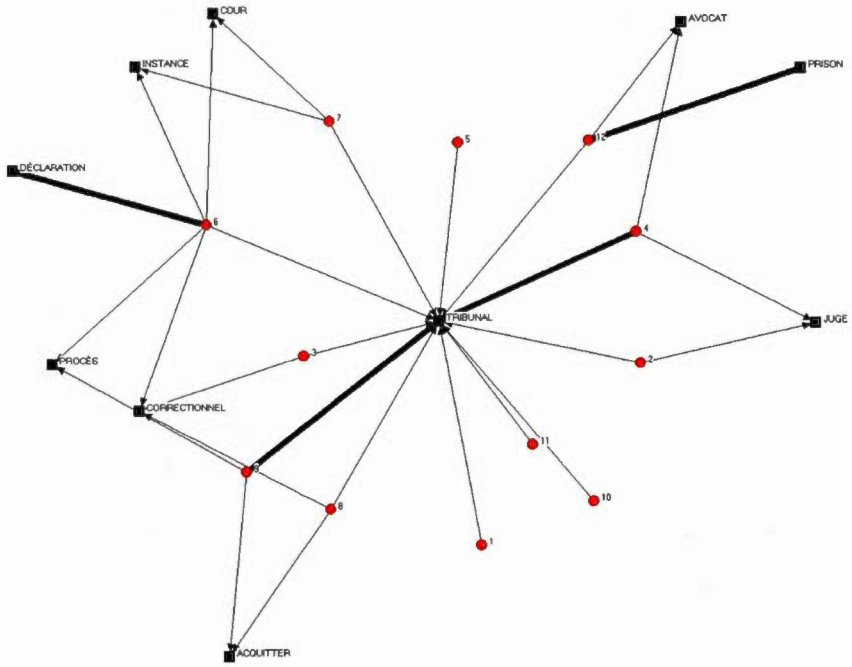
Représentation graphique du lexique de la classe 37.



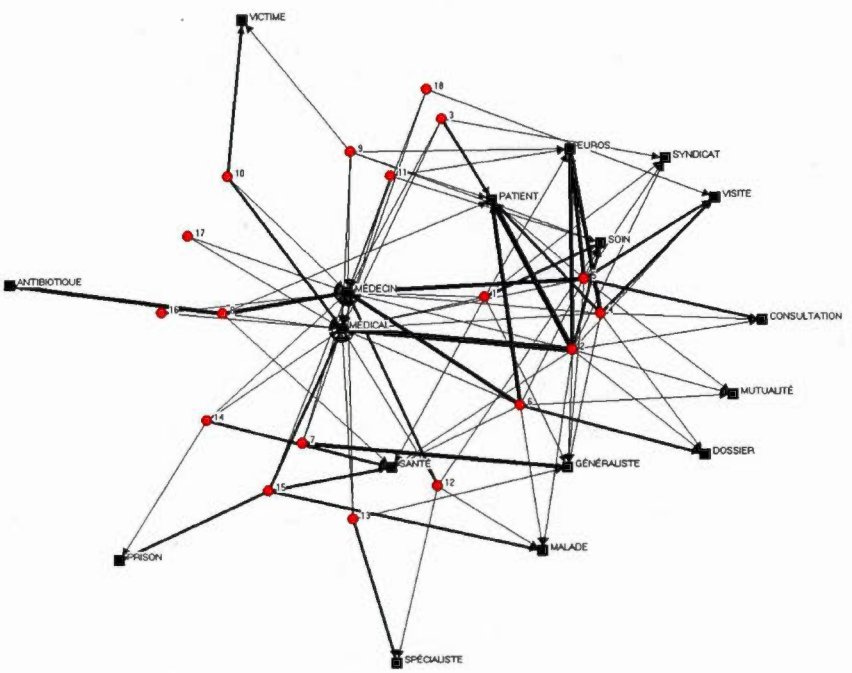
Représentation graphique du lexique de la classe 38.



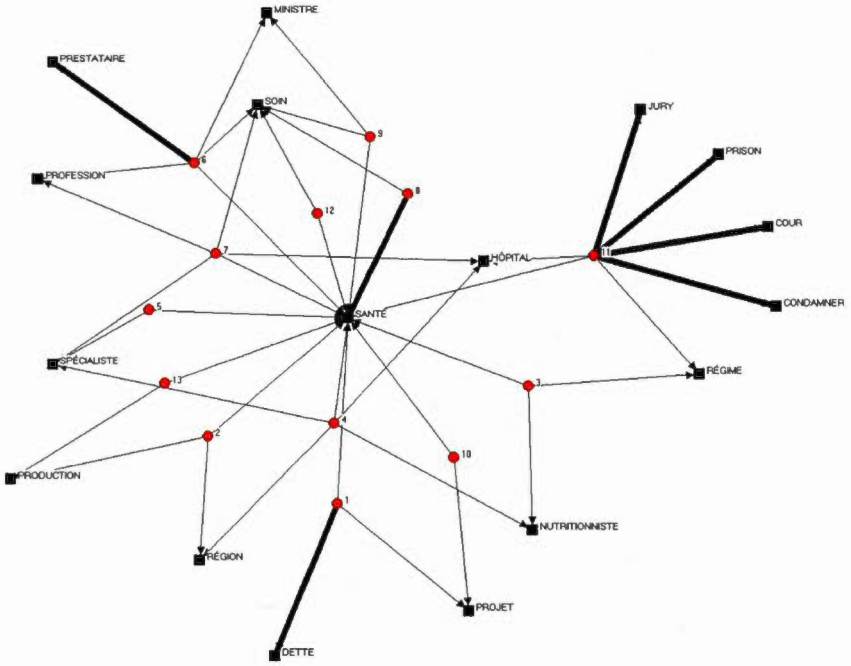
Représentation graphique du lexique de la classe 39.



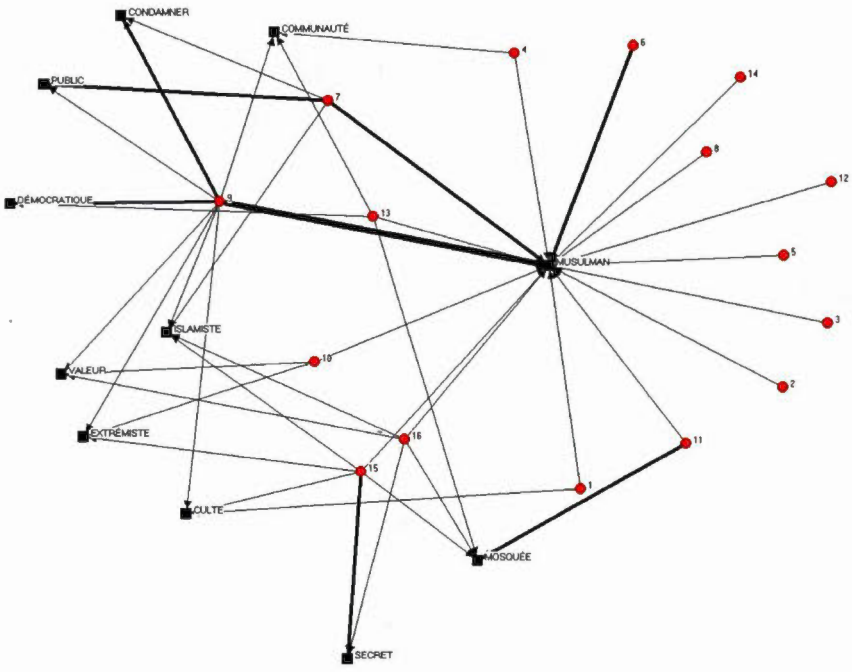
Représentation graphique du lexique de la classe 40.



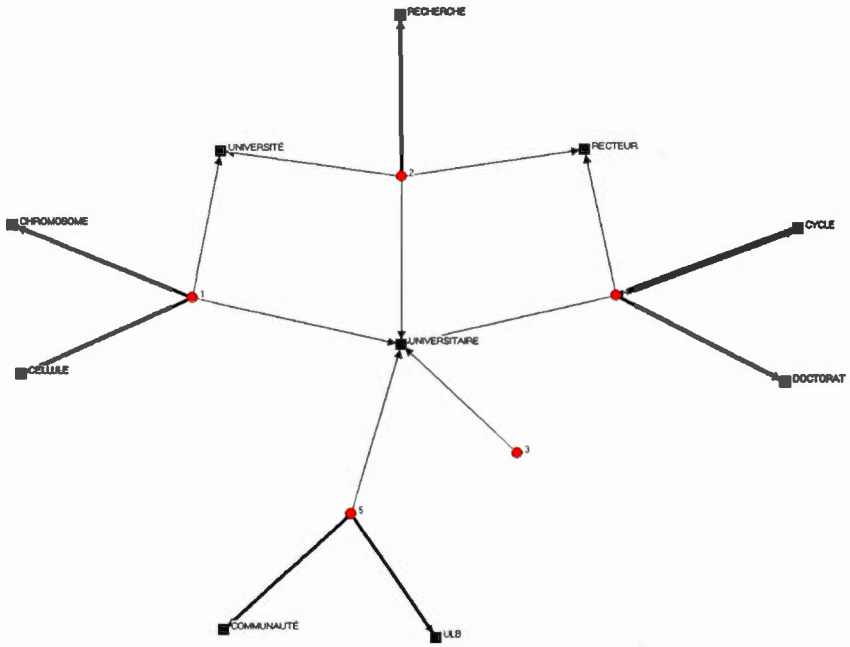
Représentation graphique du lexique de la classe 41.



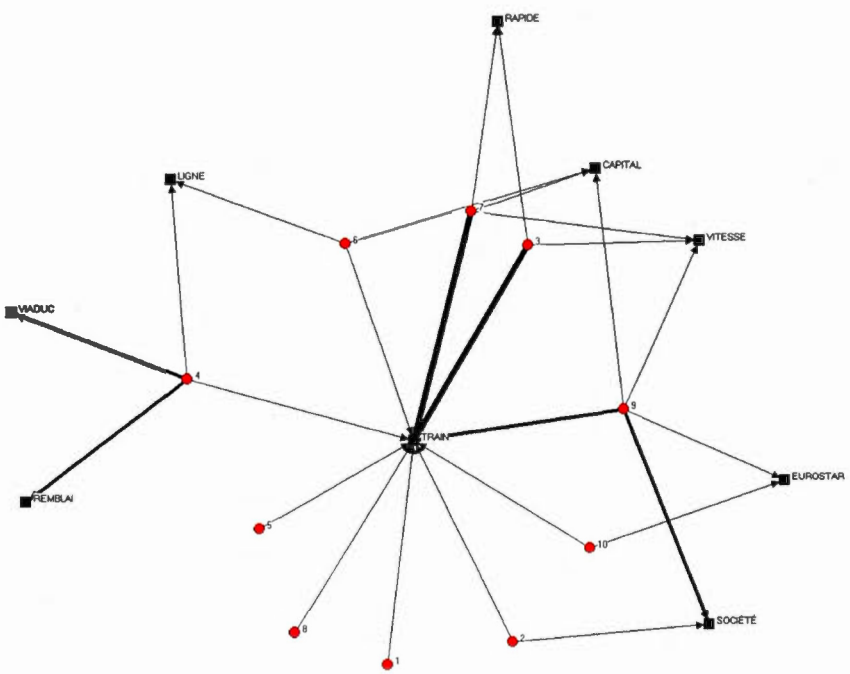
Représentation graphique du lexique de la classe 42.



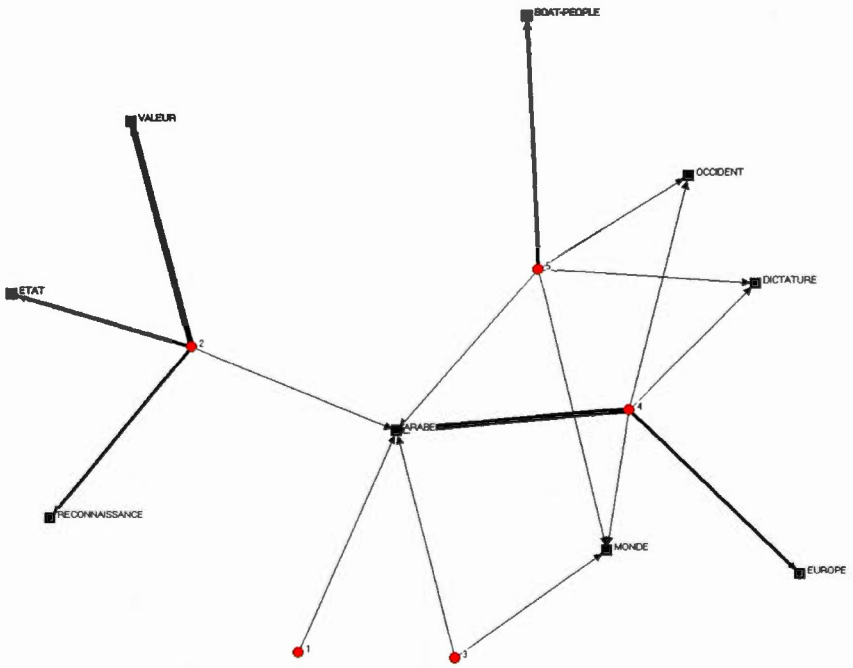
Représentation graphique du lexique de la classe 43.



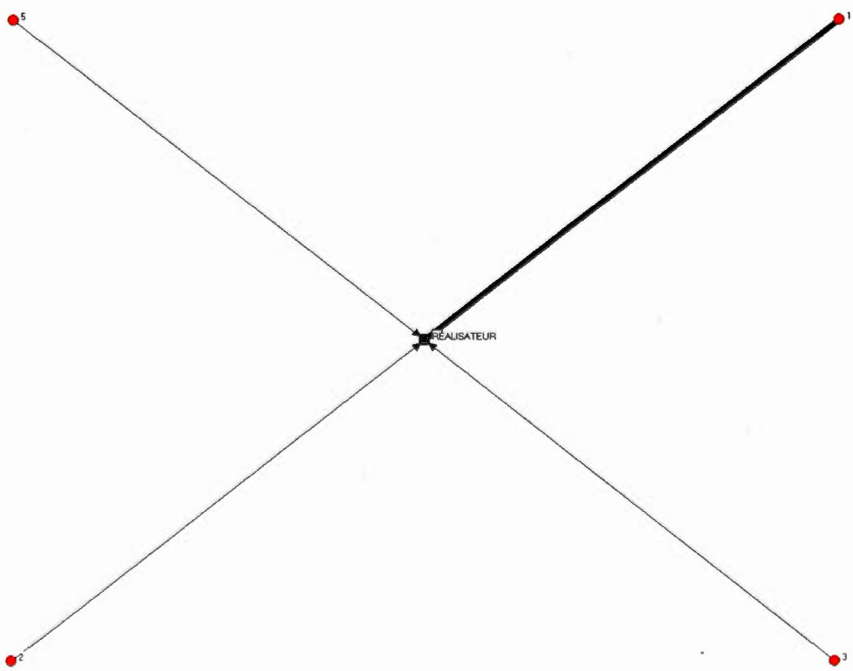
Représentation graphique du lexique de la classe 44.



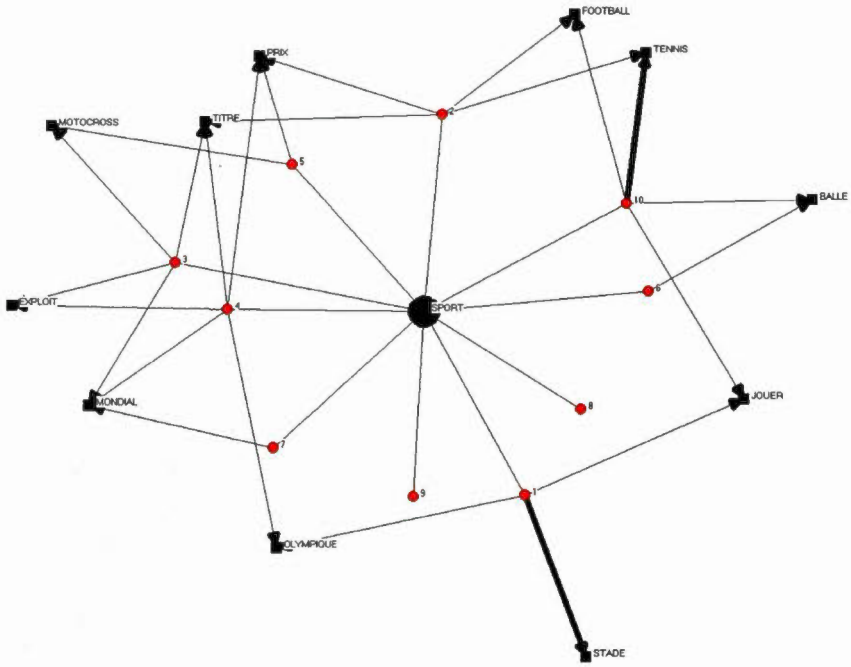
Représentation graphique du lexique de la classe 45.



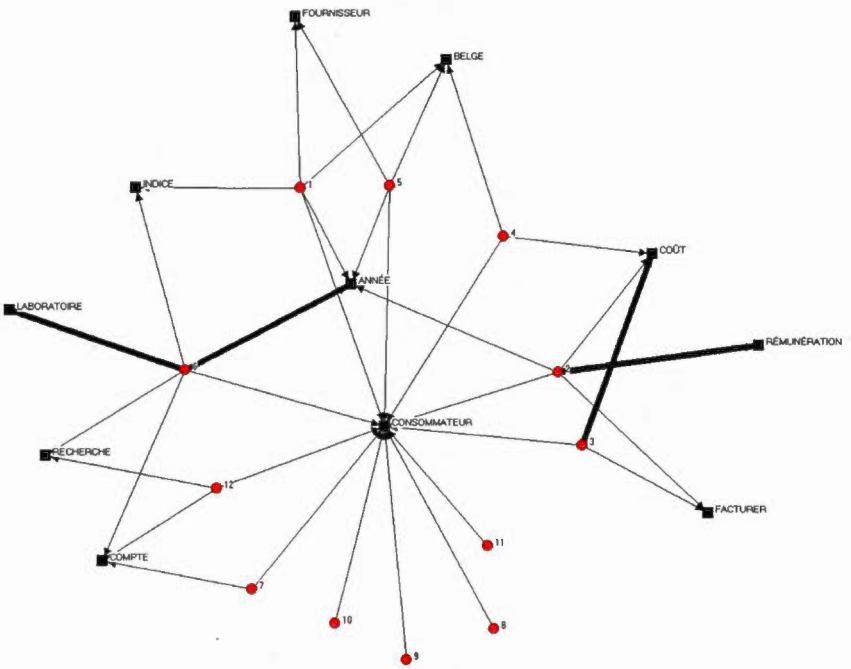
Représentation graphique du lexique de la classe 46.



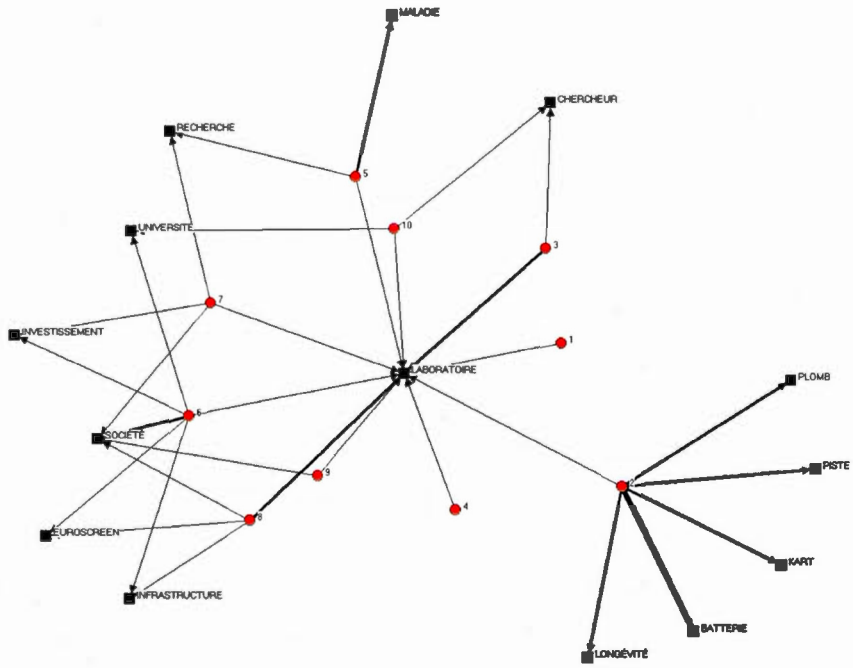
Représentation graphique du lexique de la classe 47.



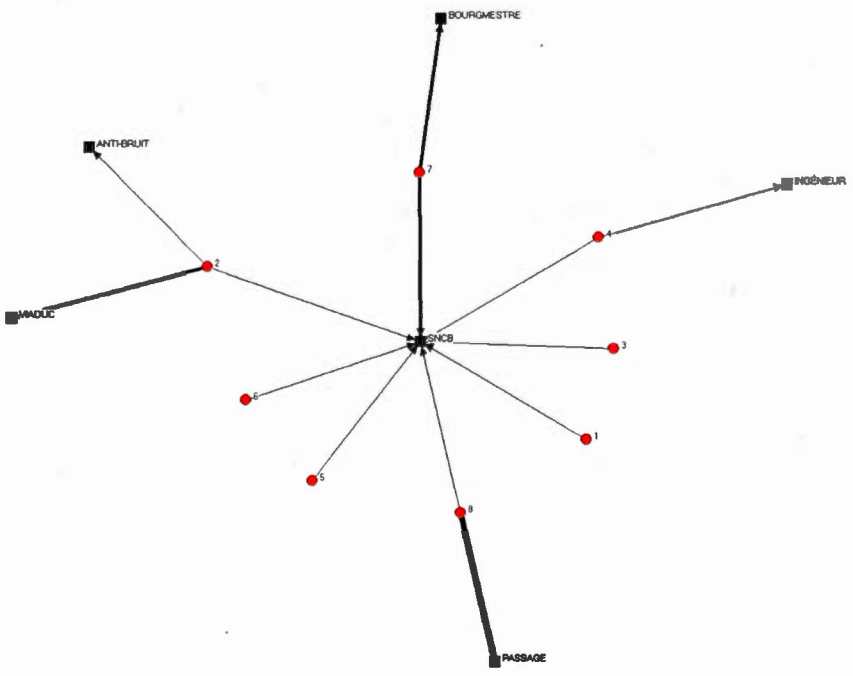
Représentation graphique du lexique de la classe 48.



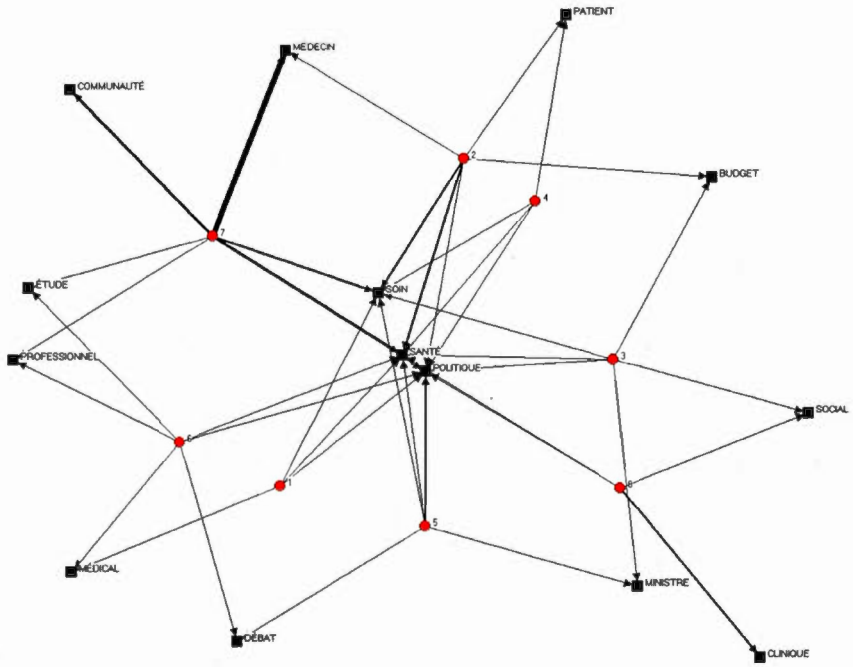
Représentation graphique du lexique de la classe 49.



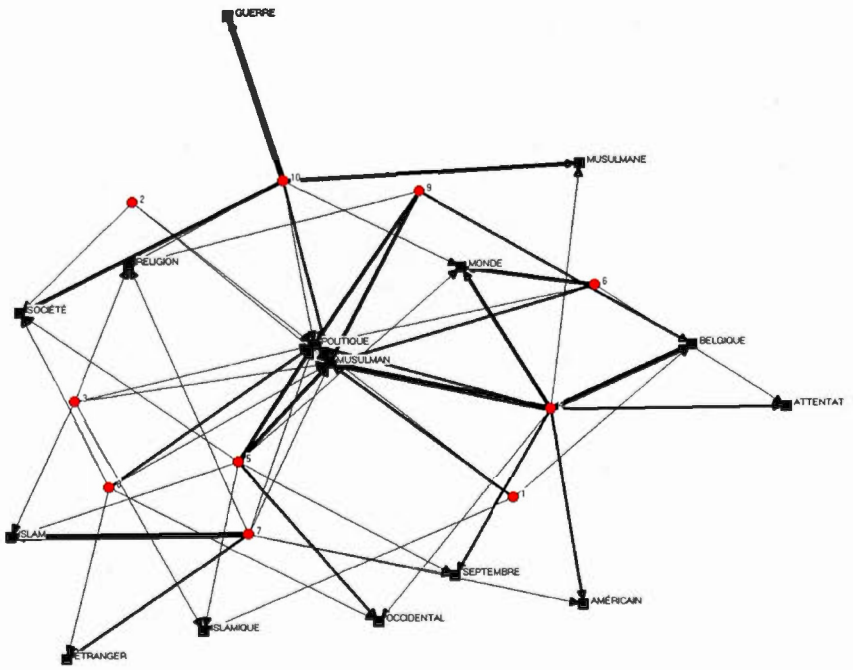
Représentation graphique du lexique de la classe 50.



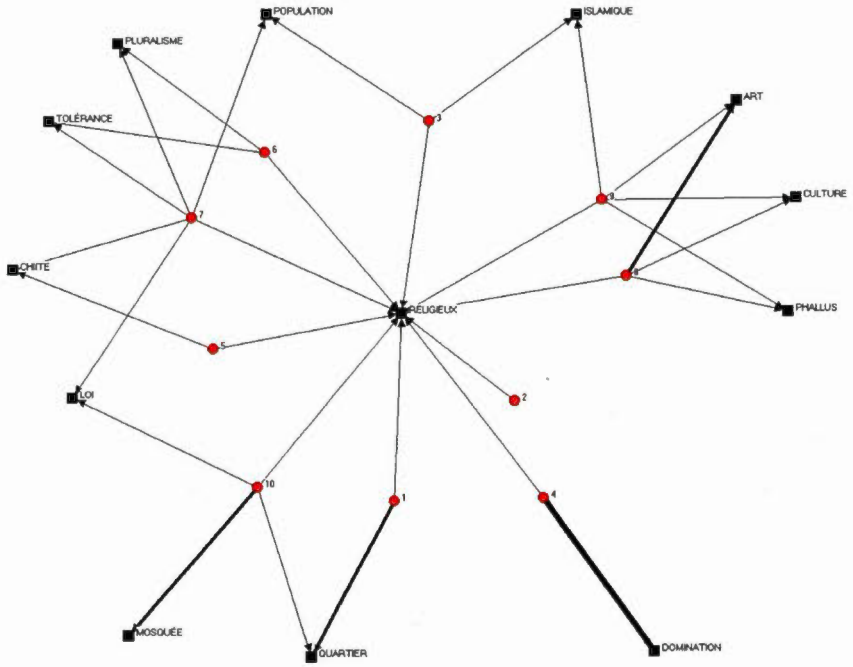
Représentation graphique du lexique de la classe 51.



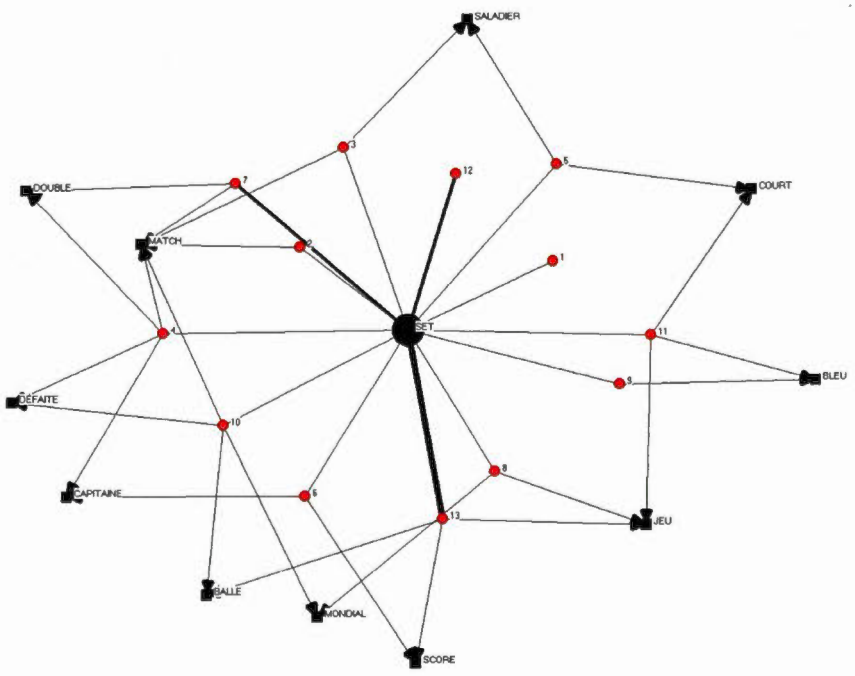
Représentation graphique du lexique de la classe 52.



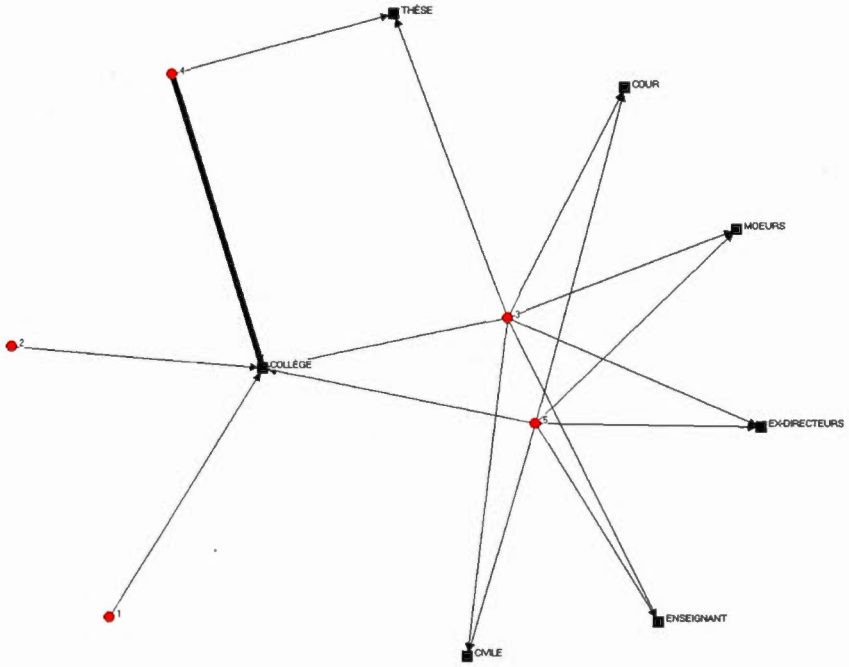
Représentation graphique du lexique de la classe 53.



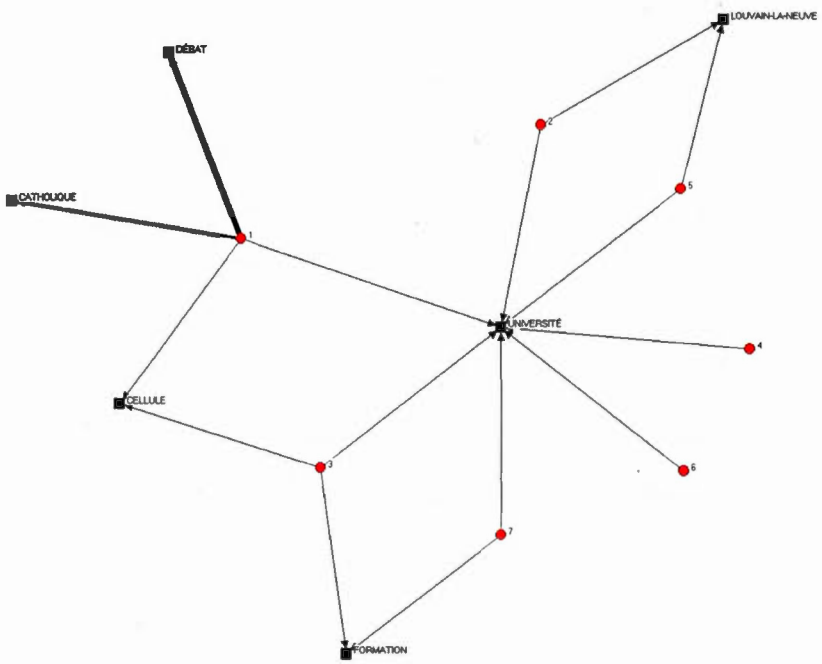
Représentation graphique du lexique de la classe 54.



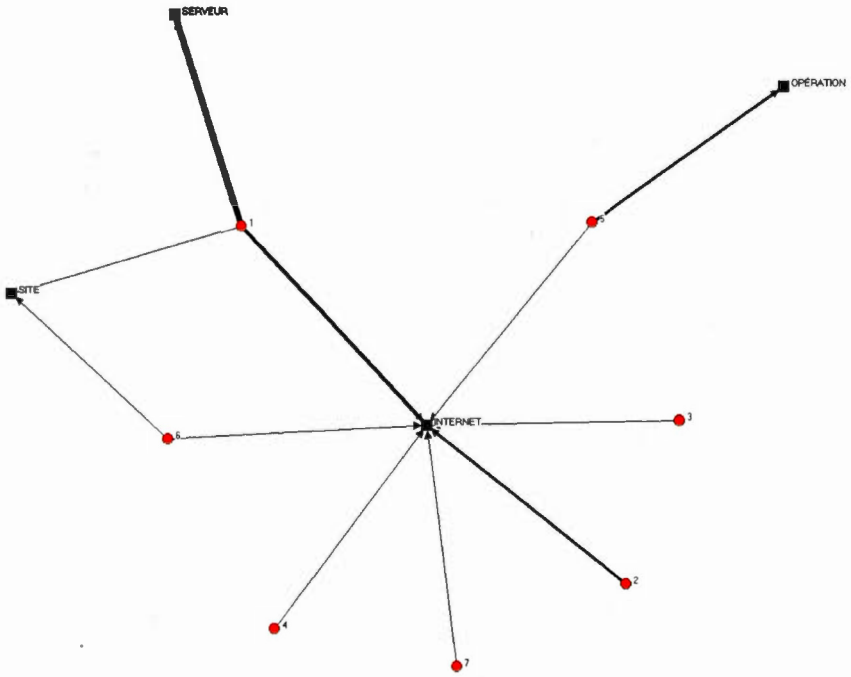
Représentation graphique du lexique de la classe 55.



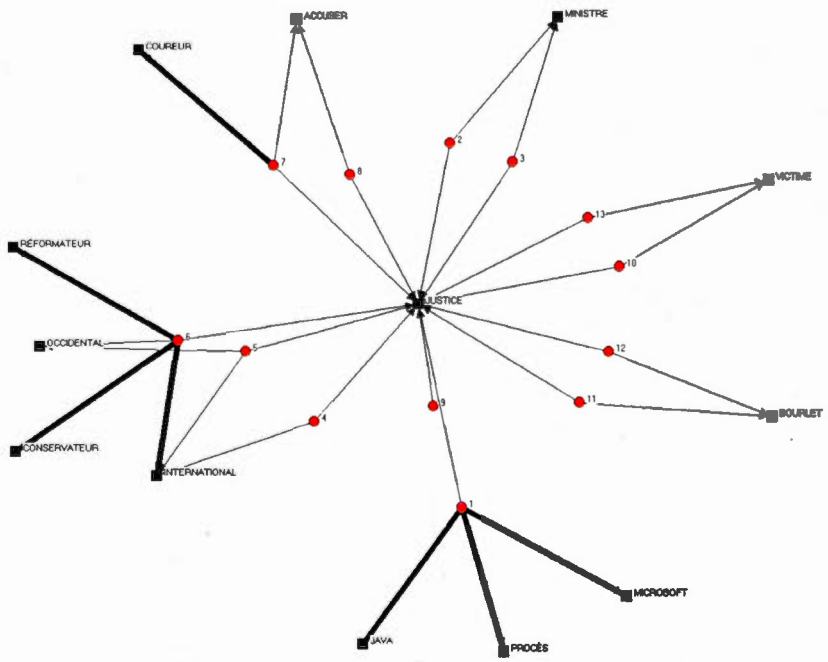
Représentation graphique du lexique de la classe 56.



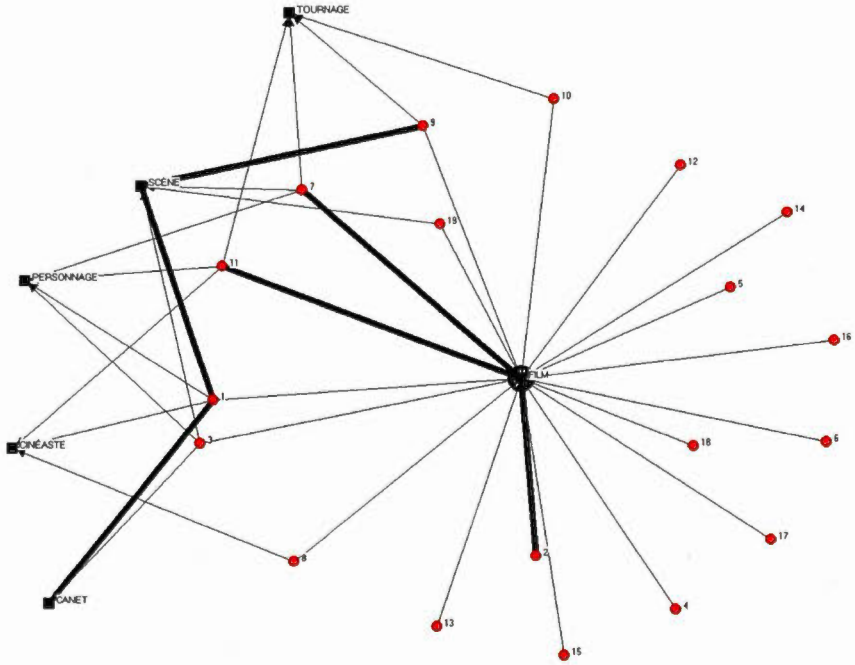
Représentation graphique du lexique de la classe 57.



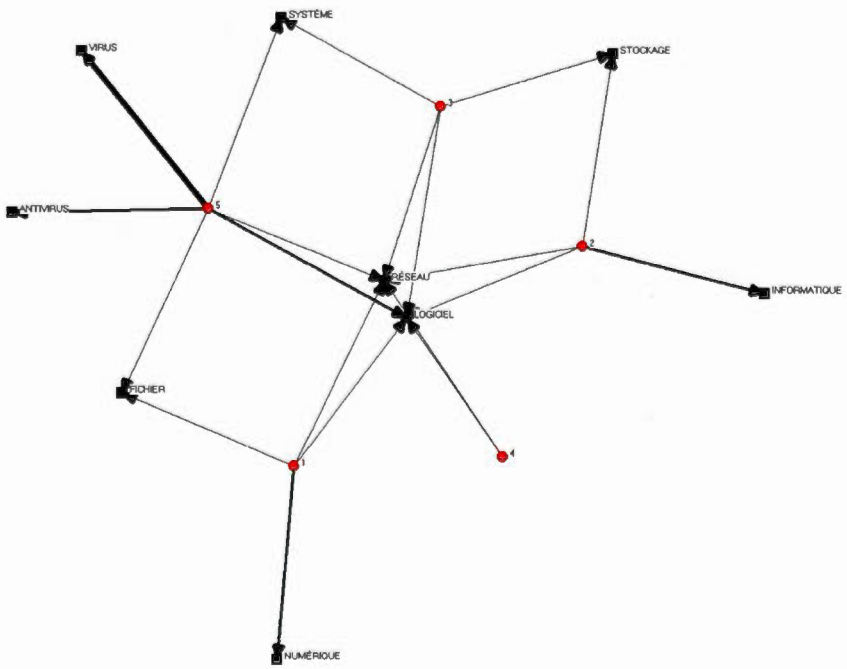
Représentation graphique du lexique de la classe 58.



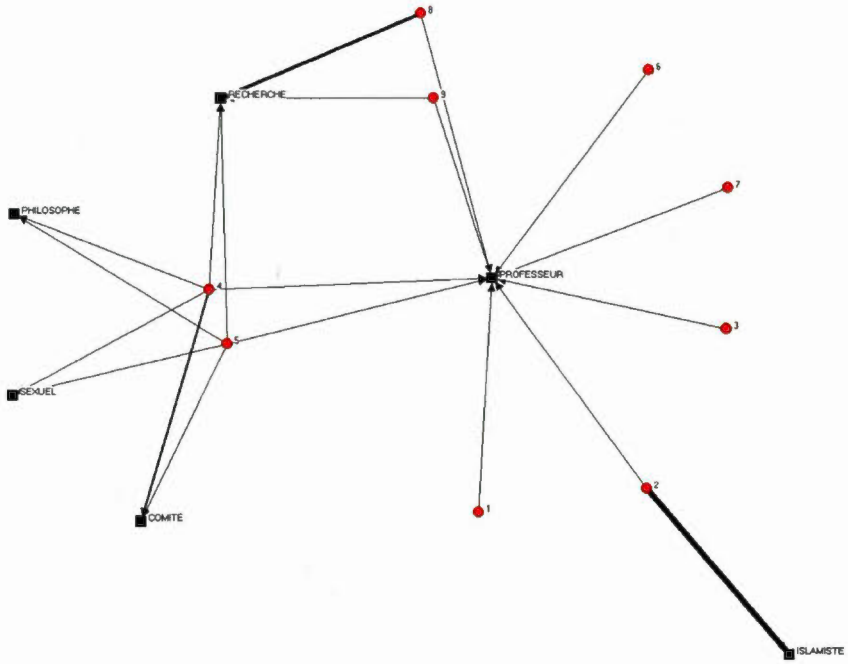
Représentation graphique du lexique de la classe 59.



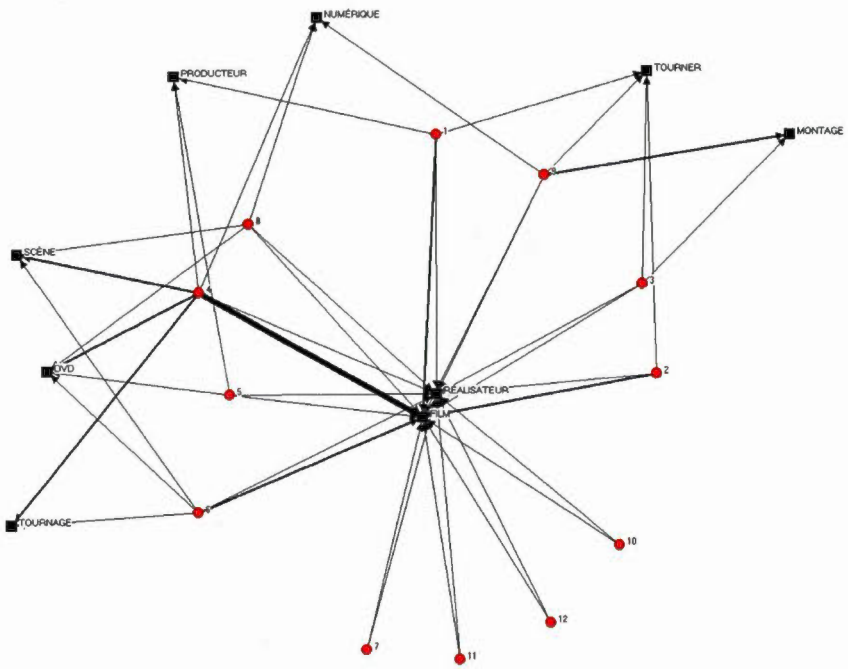
Représentation graphique du lexique de la classe 60.



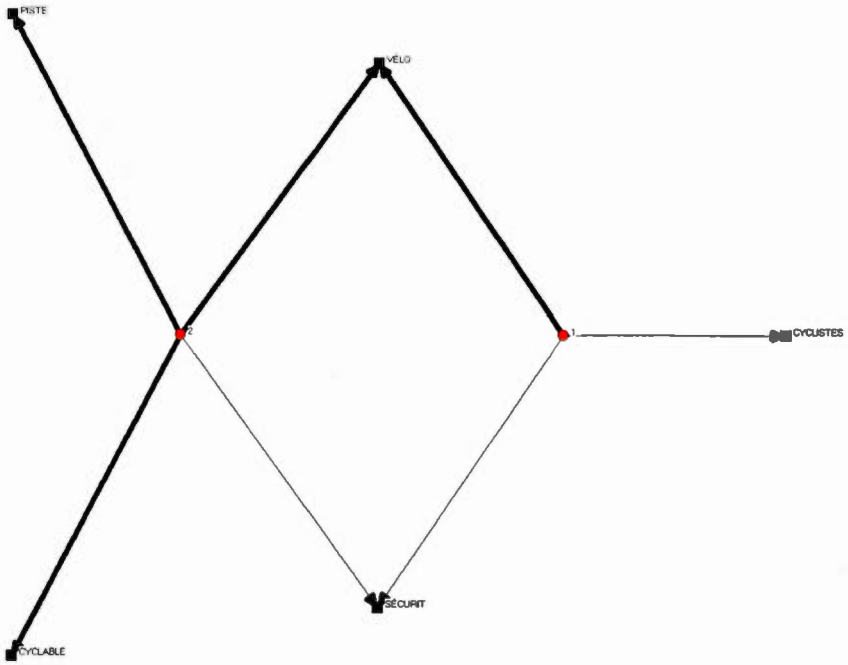
Représentation graphique du lexique de la classe 61.



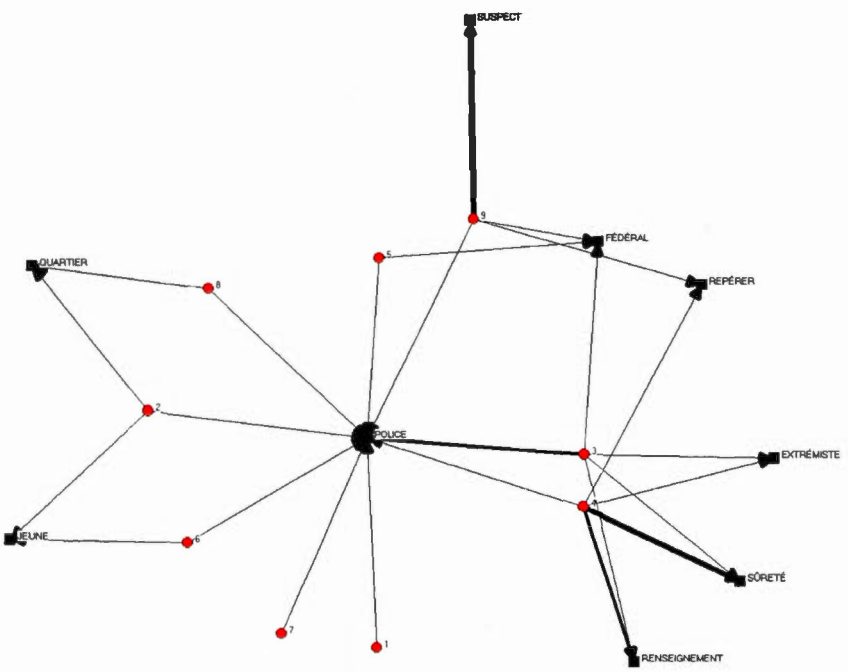
Représentation graphique du lexique de la classe 62.



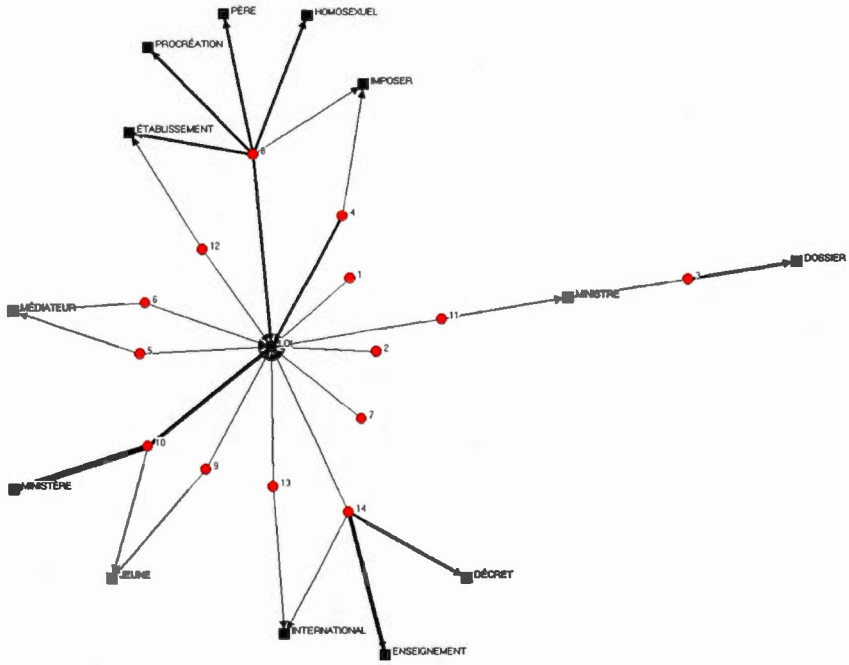
Représentation graphique du lexique de la classe 63.



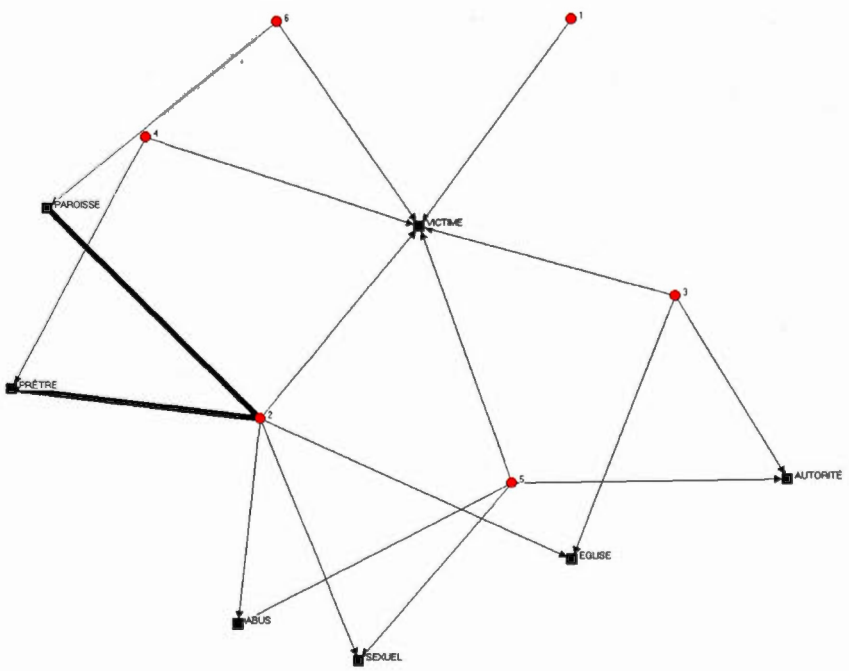
Représentation graphique du lexique de la classe 64.



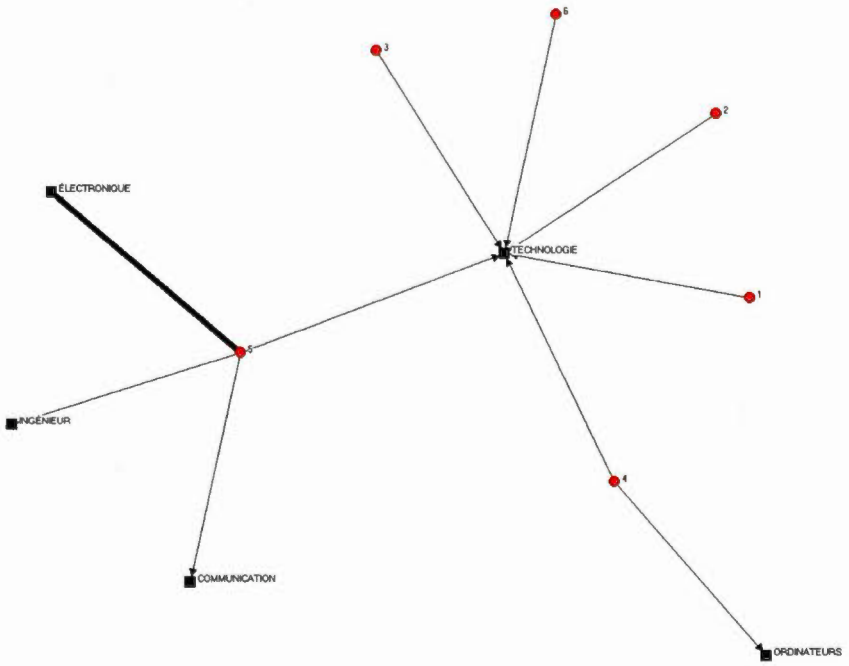
Représentation graphique du lexique de la classe 65.



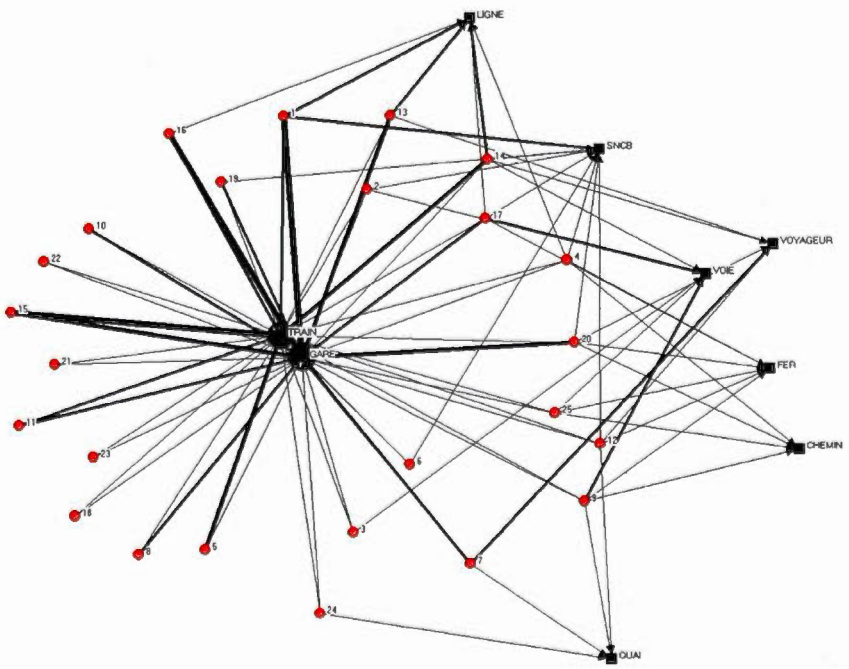
Représentation graphique du lexique de la classe 66.



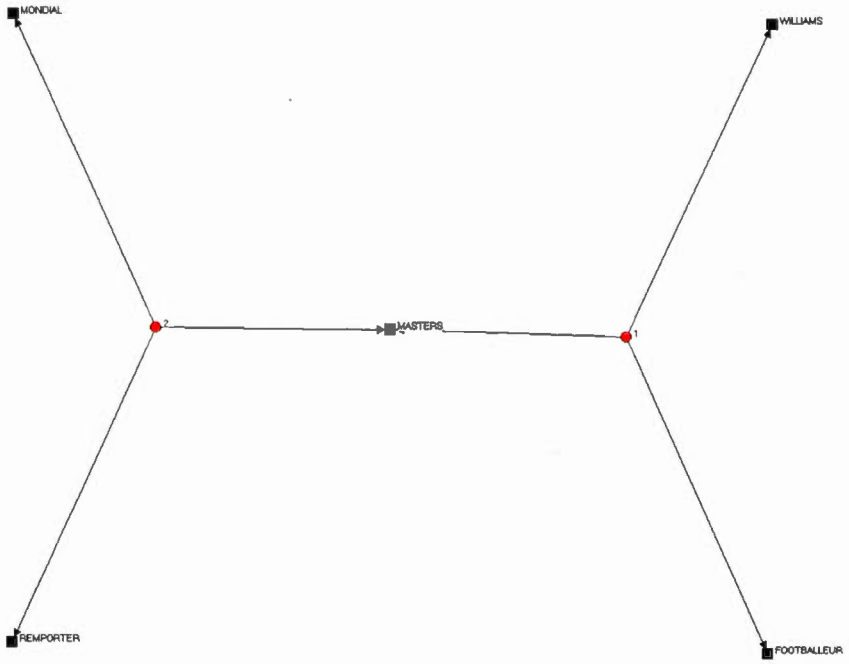
Représentation graphique du lexique de la classe 67.



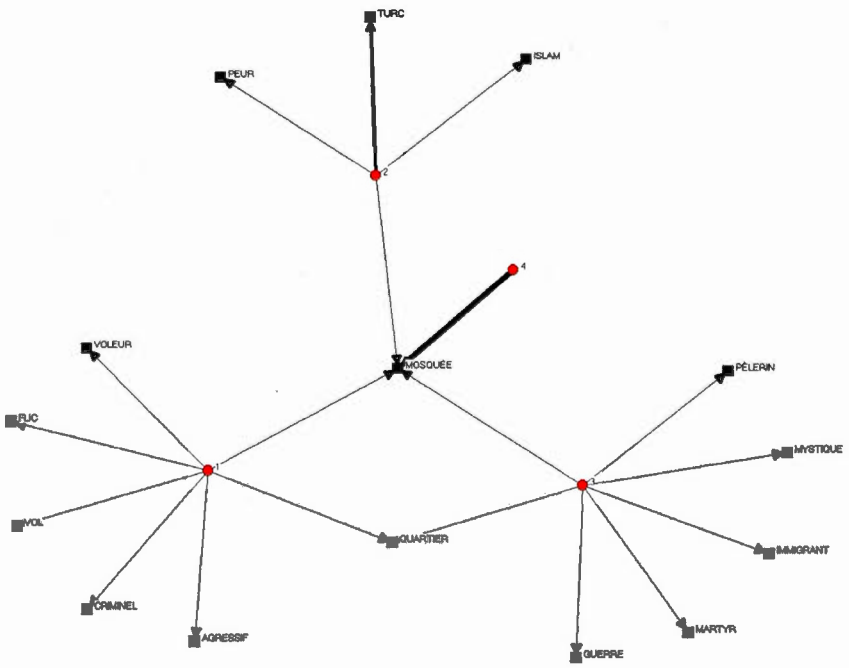
Représentation graphique du lexique de la classe 68.



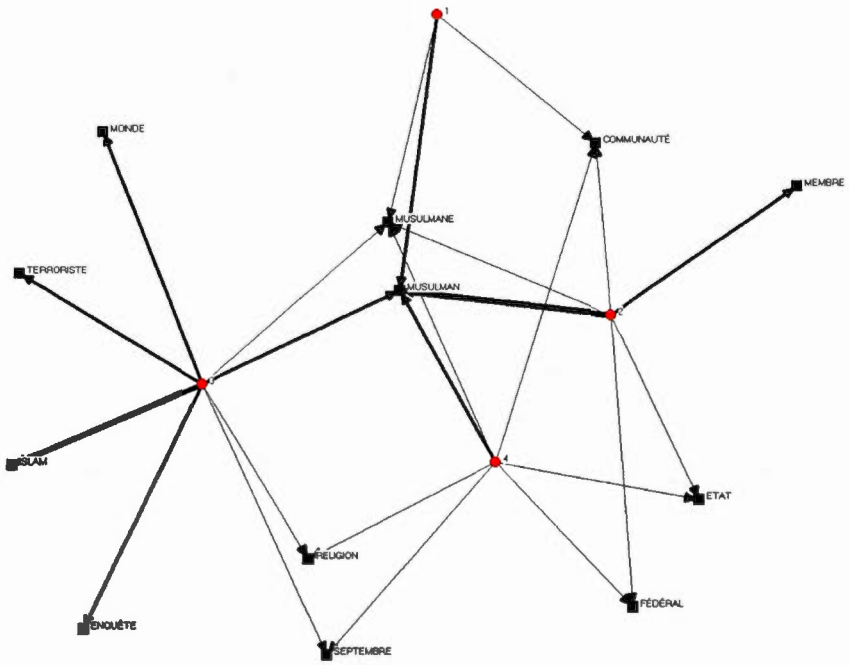
Représentation graphique du lexique de la classe 69.



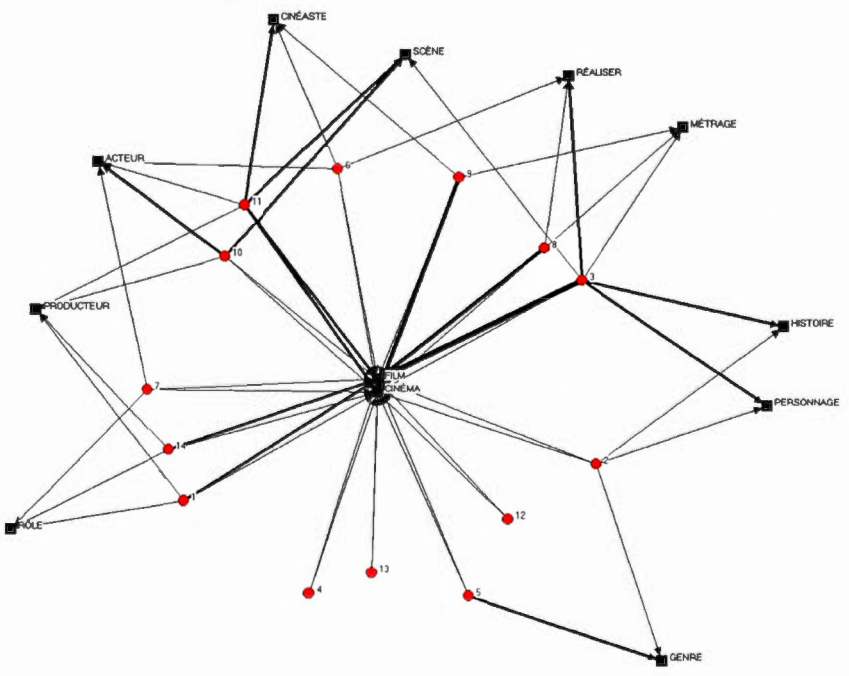
Représentation graphique du lexique de la classe 70.



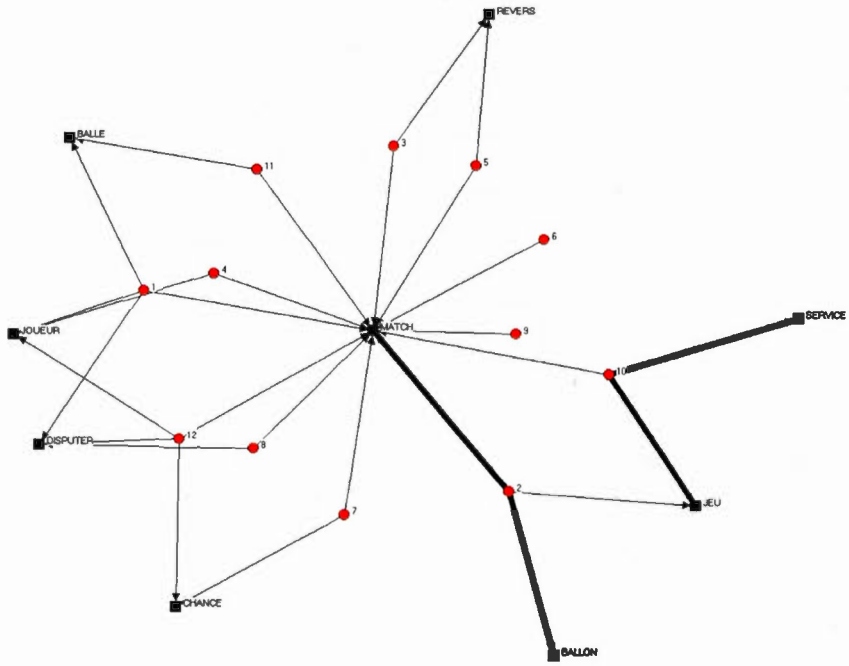
Représentation graphique du lexique de la classe 71.



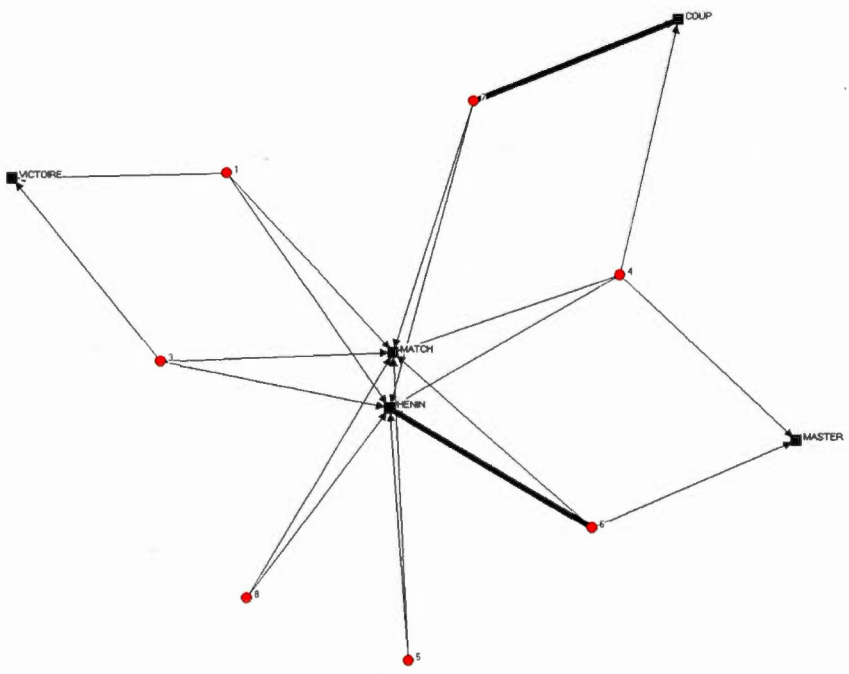
Représentation graphique du lexique de la classe 72.



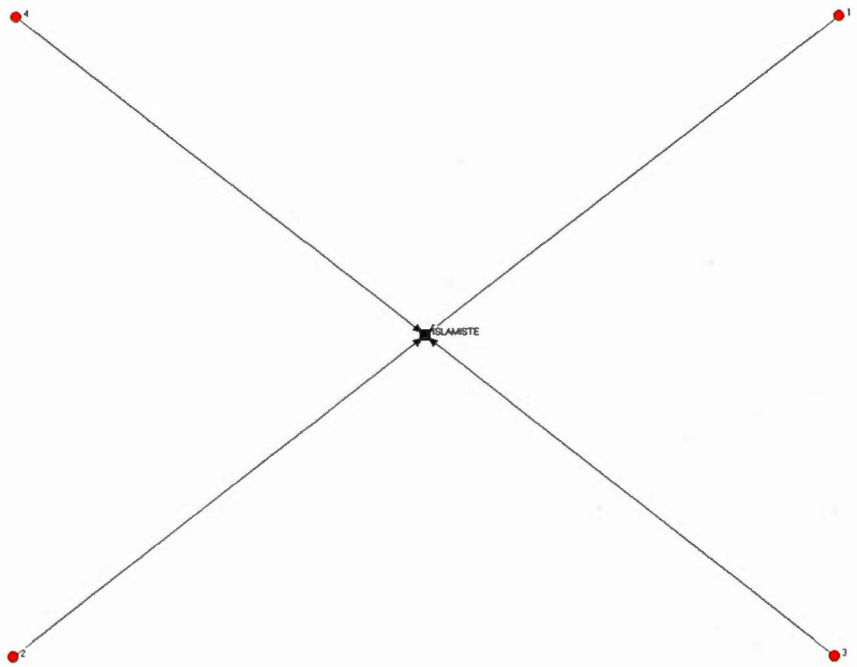
Représentation graphique du lexique de la classe 73.



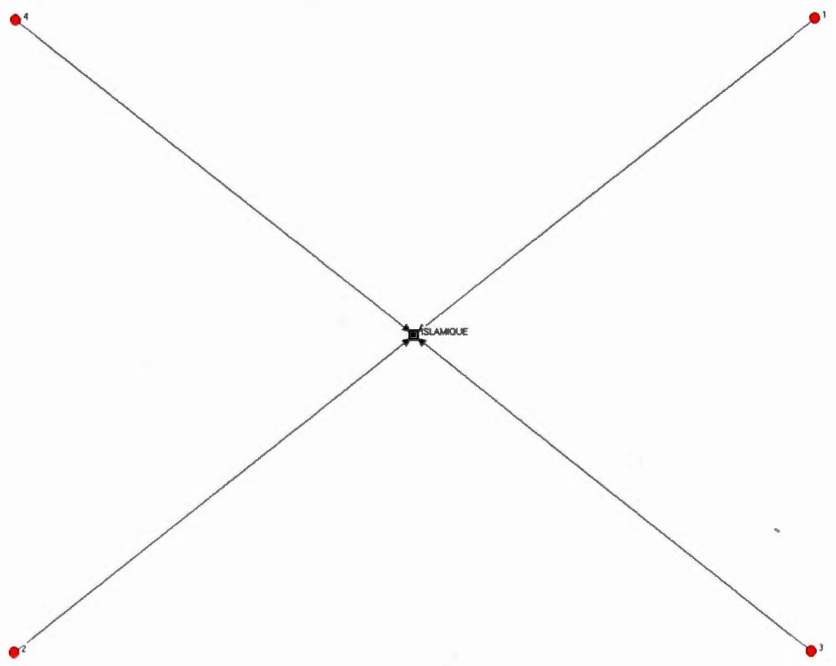
Représentation graphique du lexique de la classe 74.



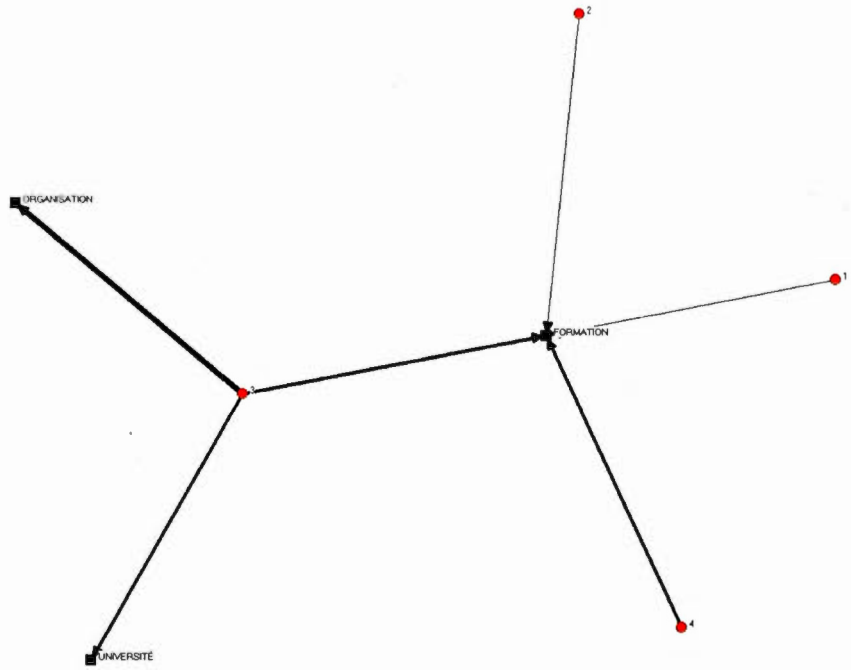
Représentation graphique du lexique de la classe 75.



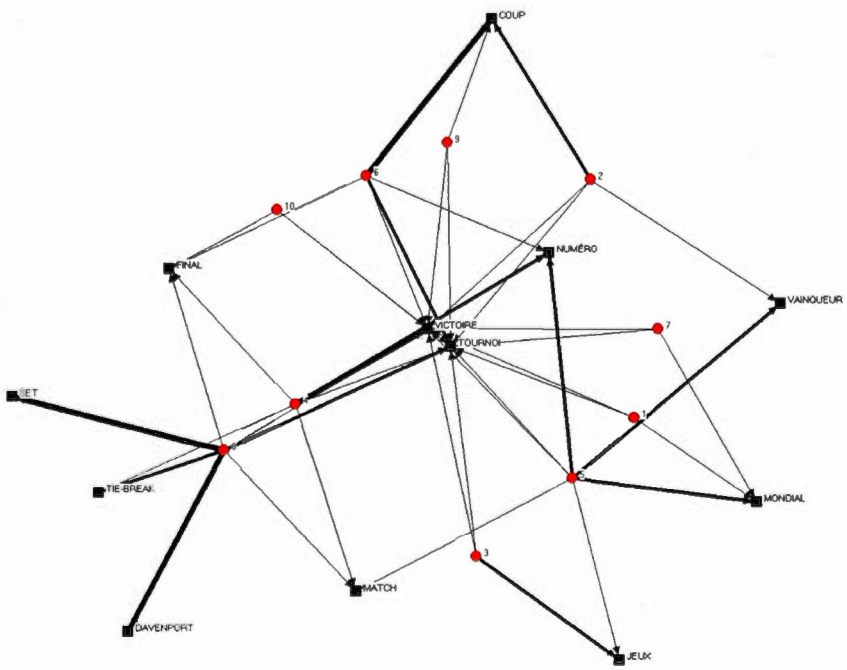
Représentation graphique du lexique de la classe 76.



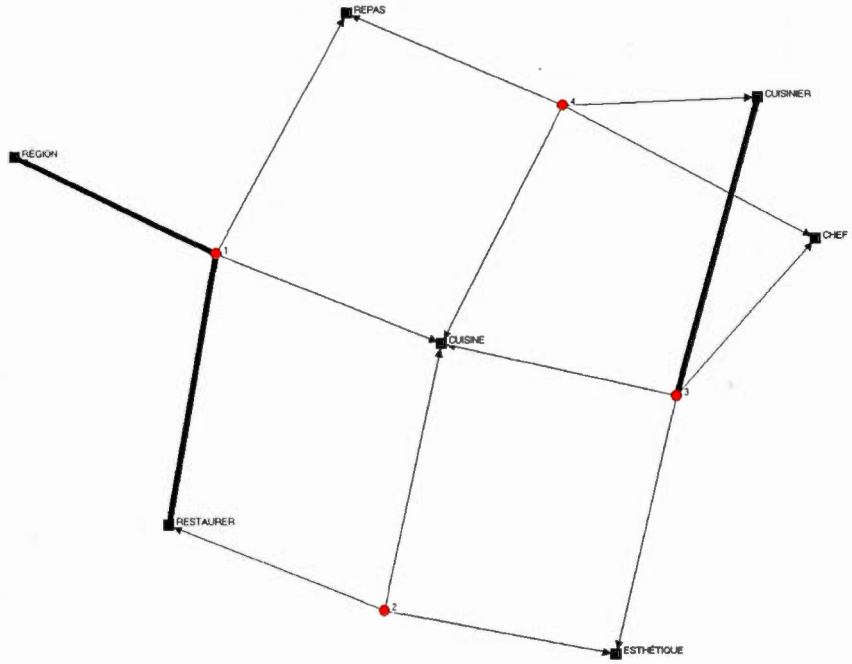
Représentation graphique du lexique de la classe 77.



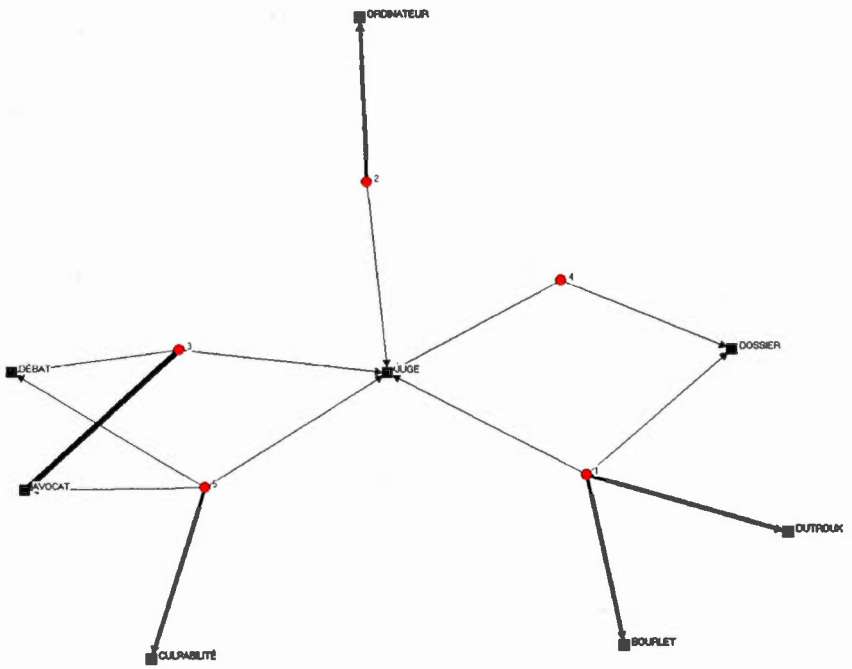
Représentation graphique du lexique de la classe 78.



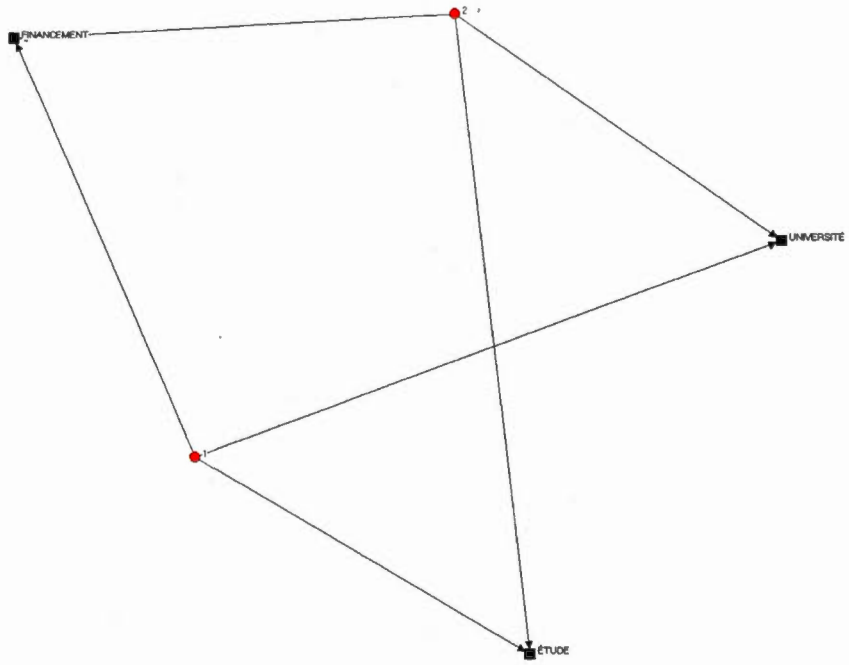
Représentation graphique du lexique de la classe 79.



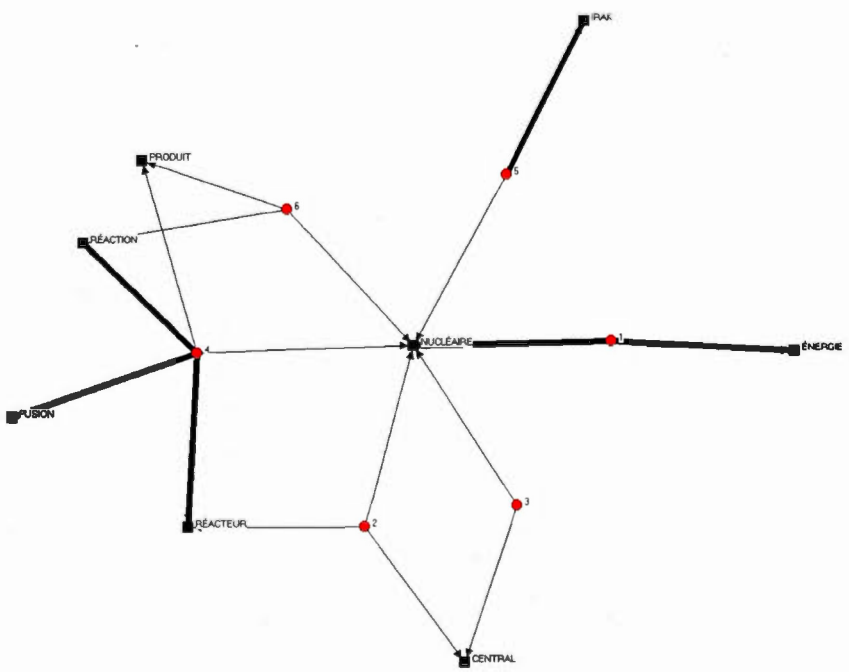
Représentation graphique du lexique de la classe 80.



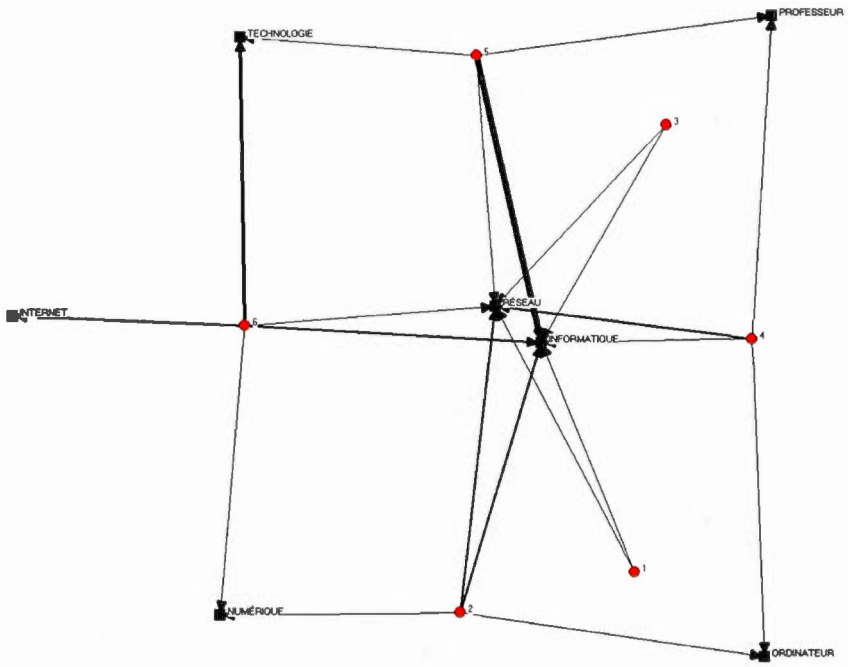
Représentation graphique du lexique de la classe 81.



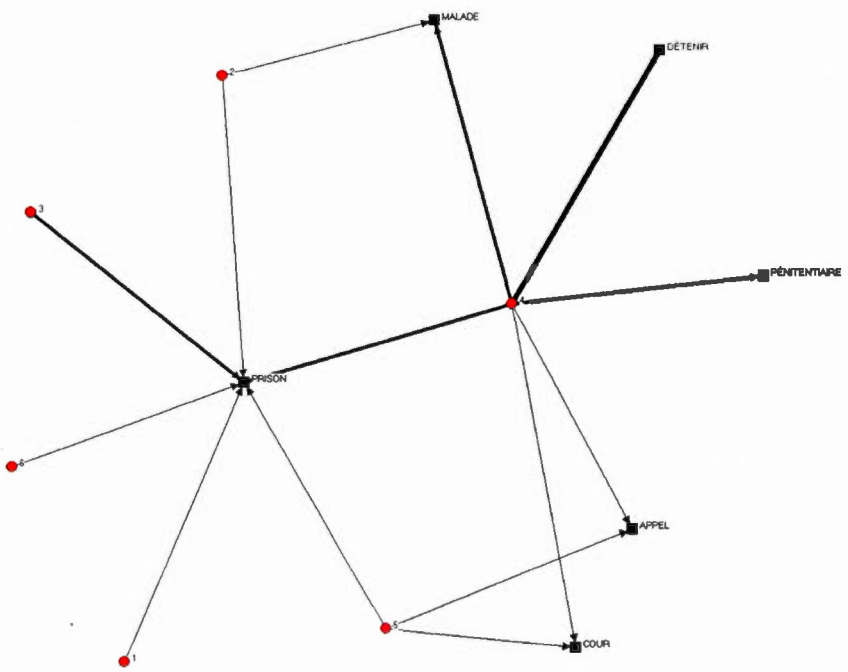
Représentation graphique du lexique de la classe 82.



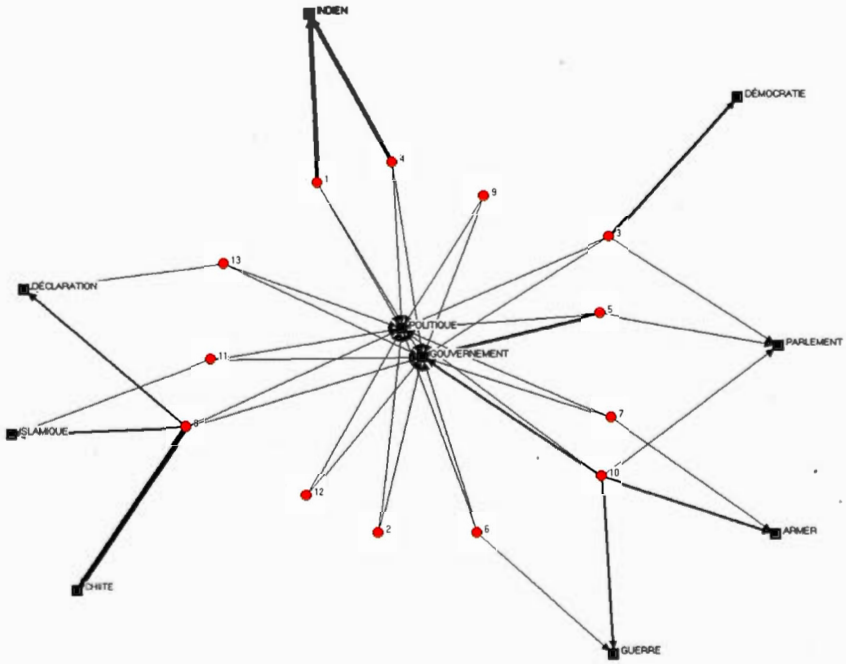
Représentation graphique du lexique de la classe 83.



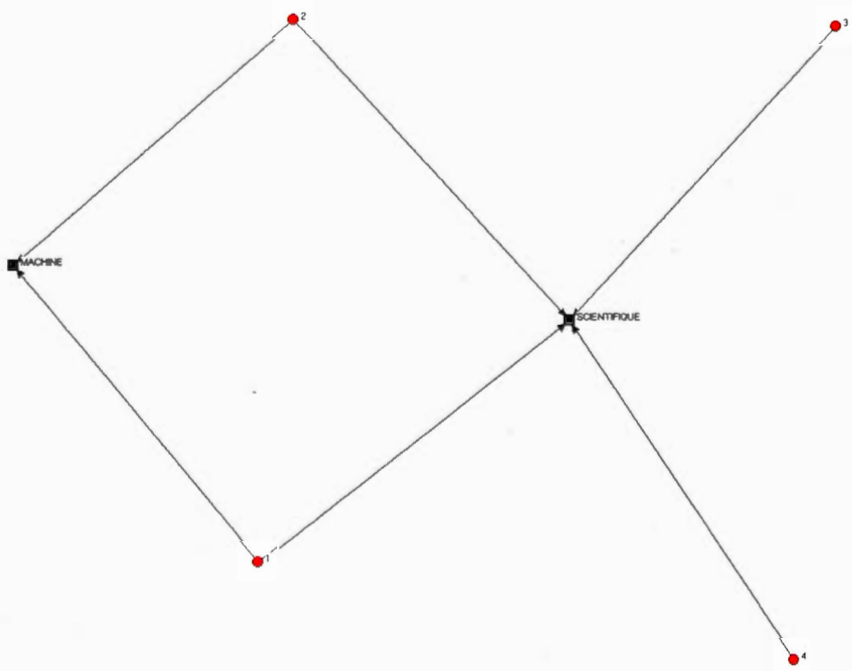
Représentation graphique du lexique de la classe 84.



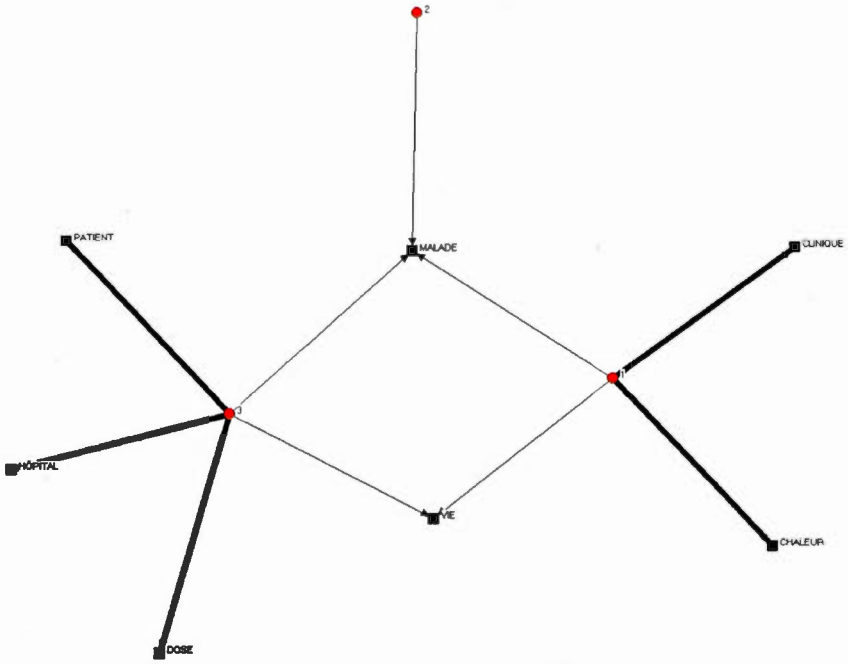
Représentation graphique du lexique de la classe 85.



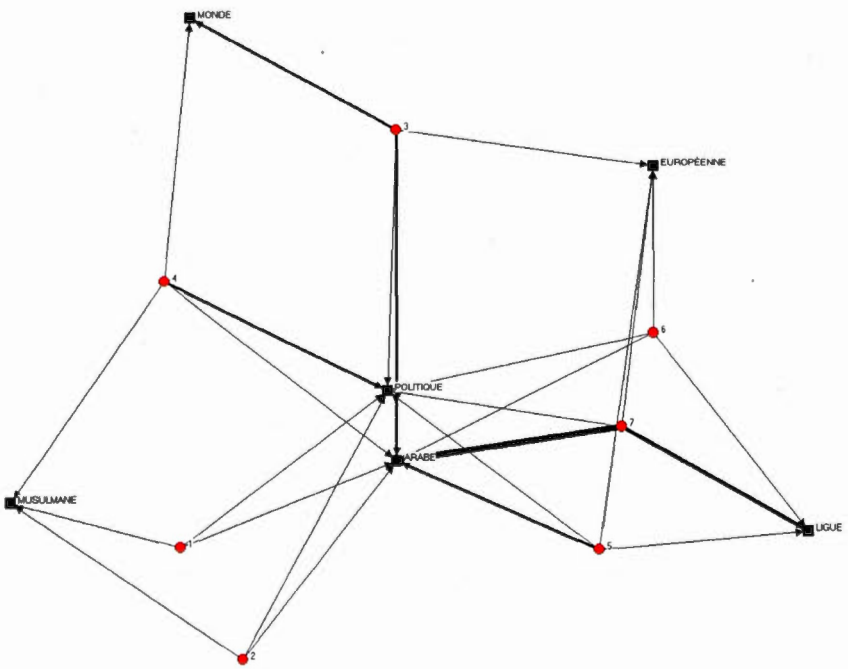
Représentation graphique du lexique de la classe 86.



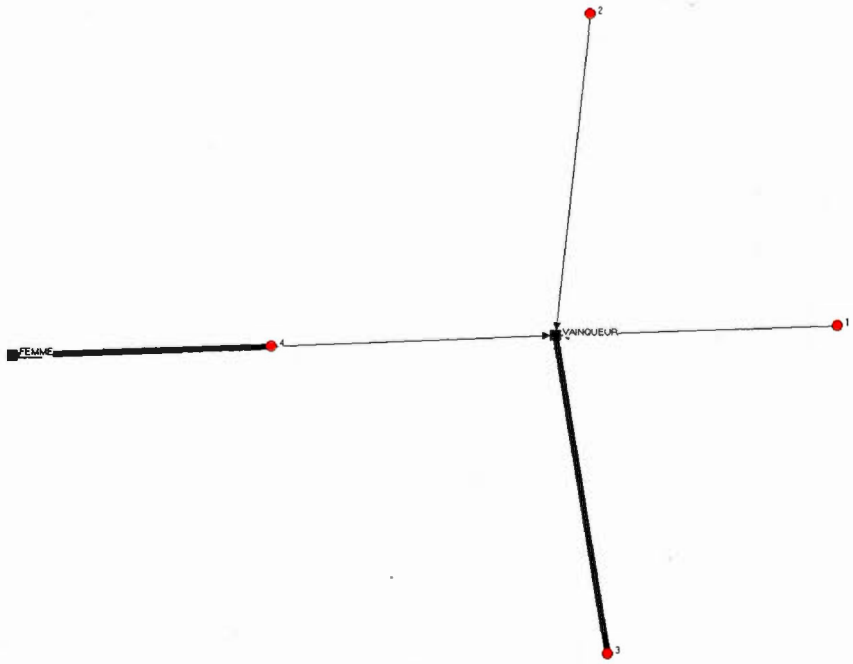
Représentation graphique du lexique de la classe 87.



Représentation graphique du lexique de la classe 88.

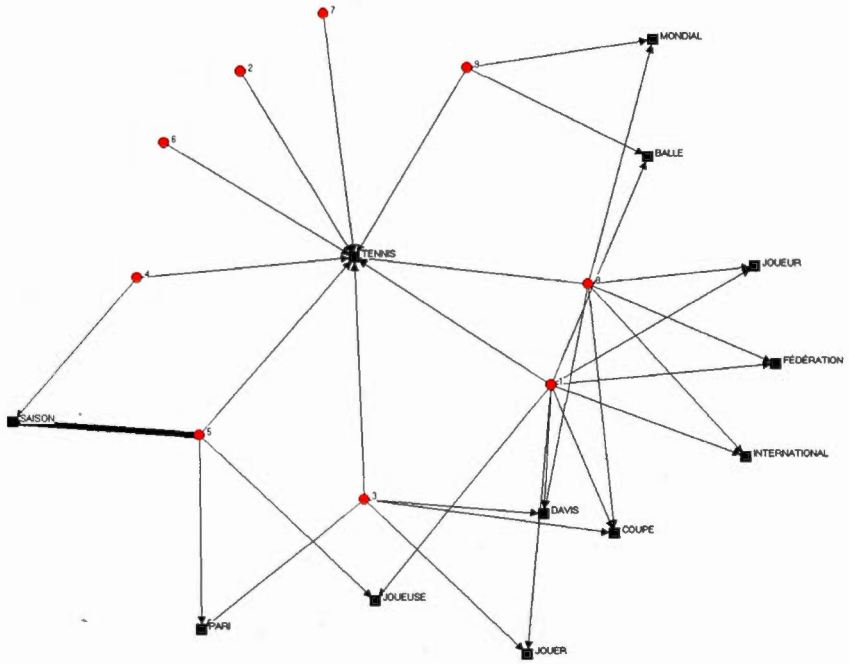


Représentation graphique du lexique de la classe 89.

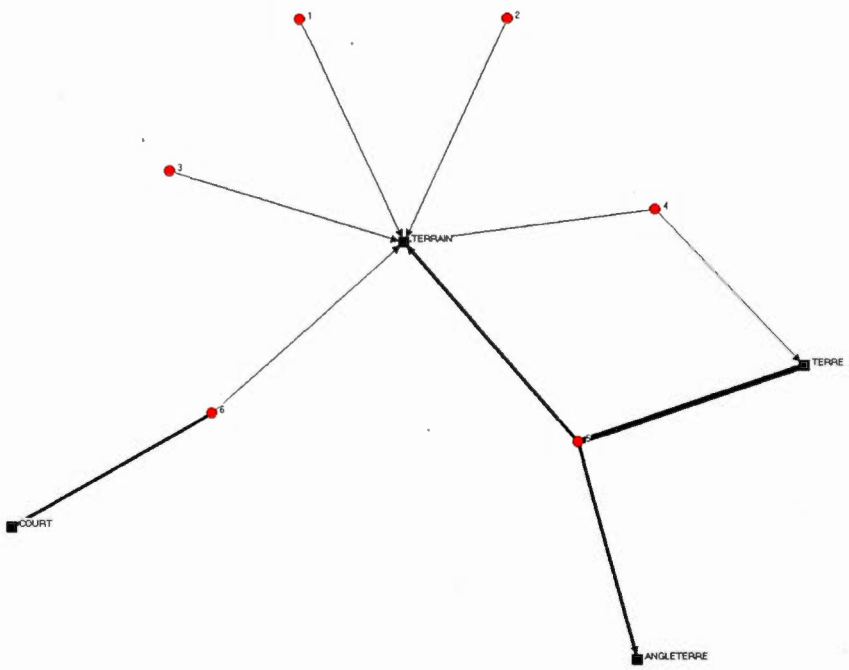


Représentation graphique du lexique de la classe 90.

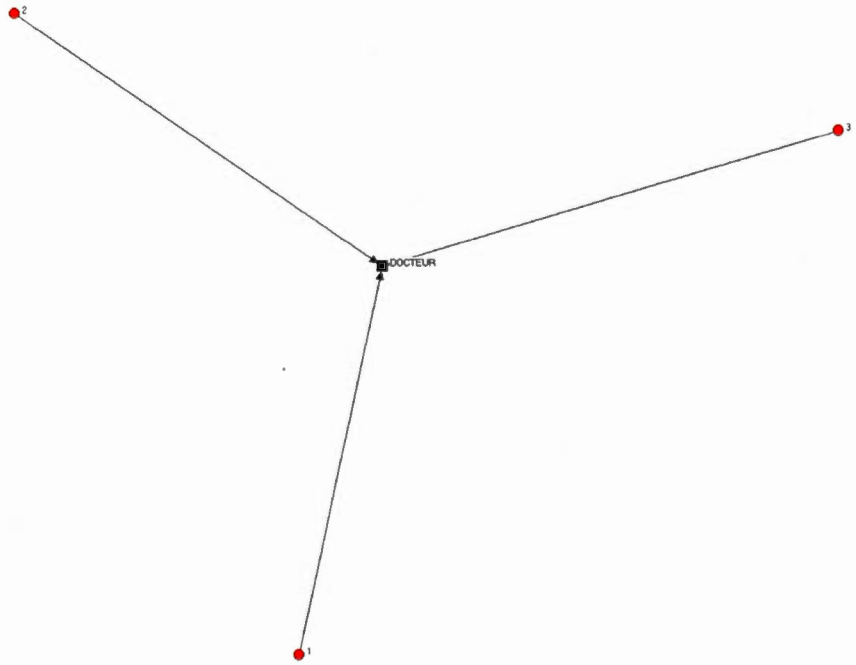
Représentation graphique du lexique de la classe 91.



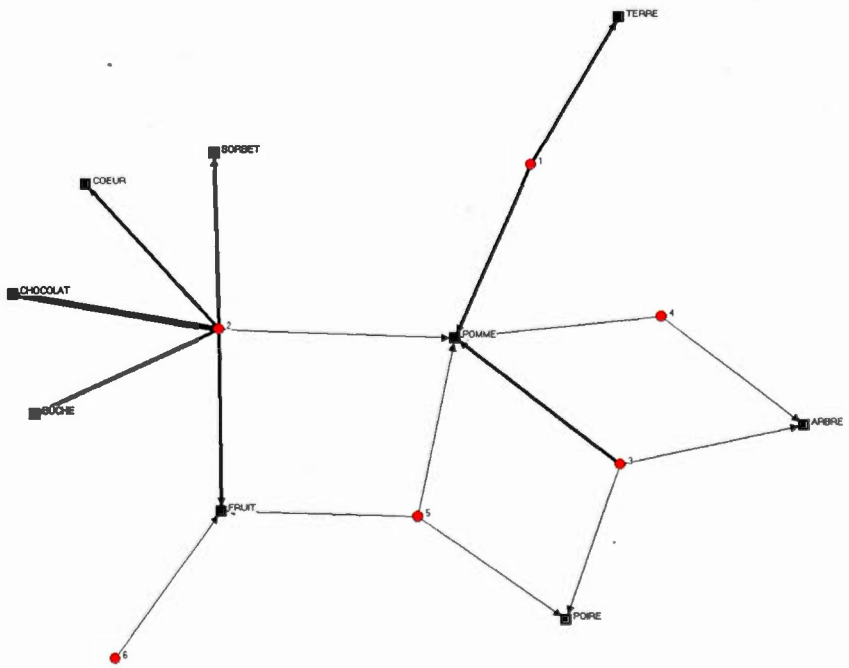
Représentation graphique du lexique de la classe 92.



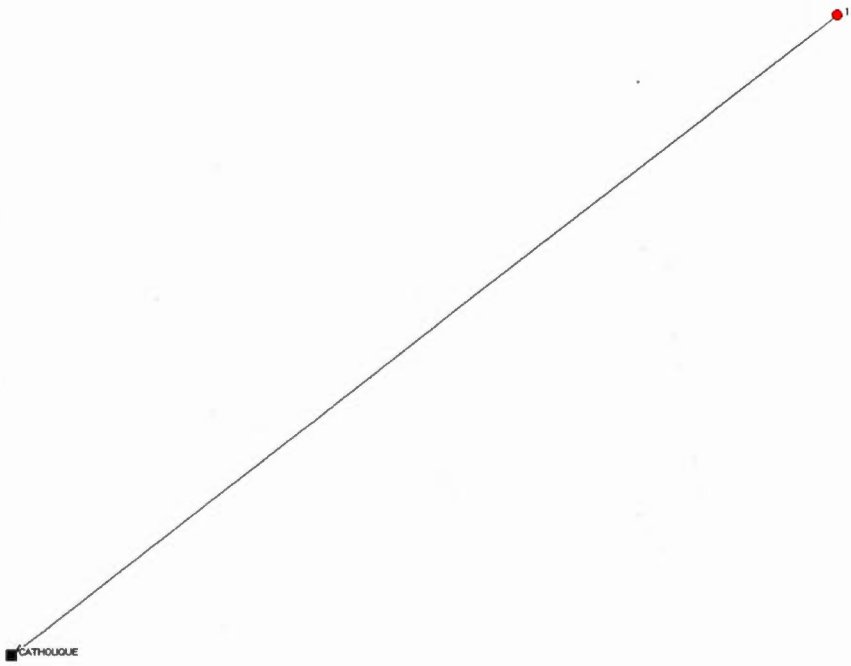
Représentation graphique du lexique de la classe 93.



Représentation graphique du lexique de la classe 94.

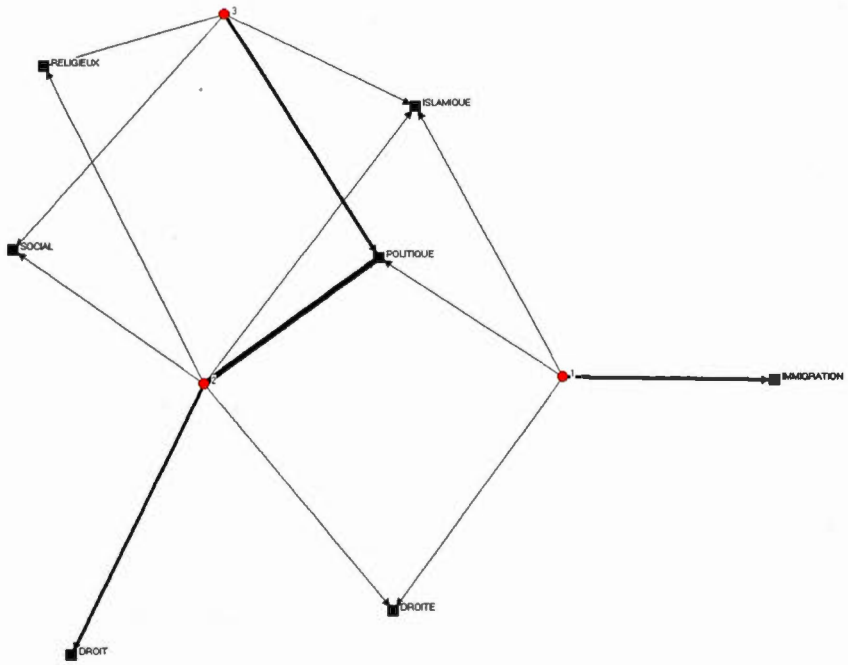


Représentation graphique du lexique de la classe 95.

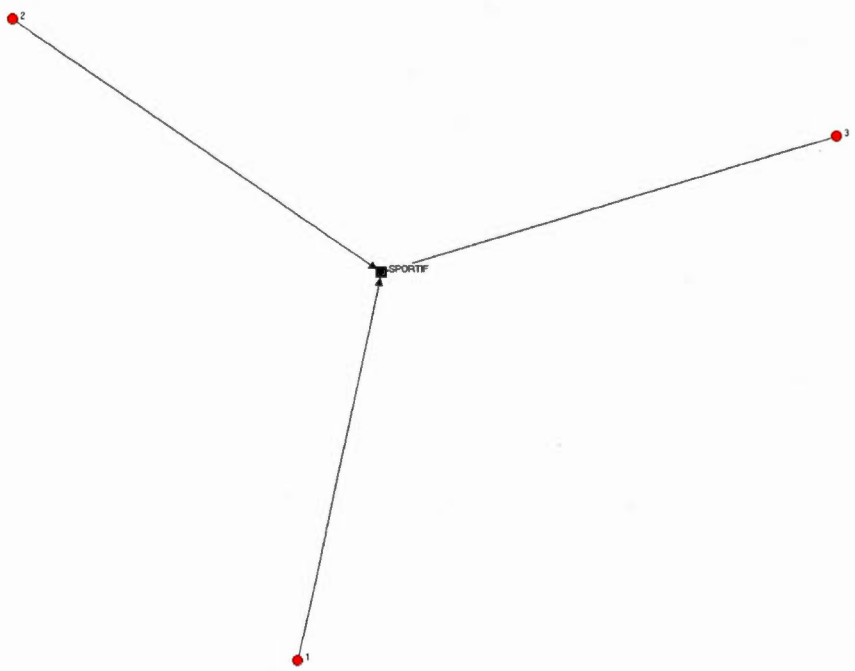


Représentation graphique du lexique de la classe 96.

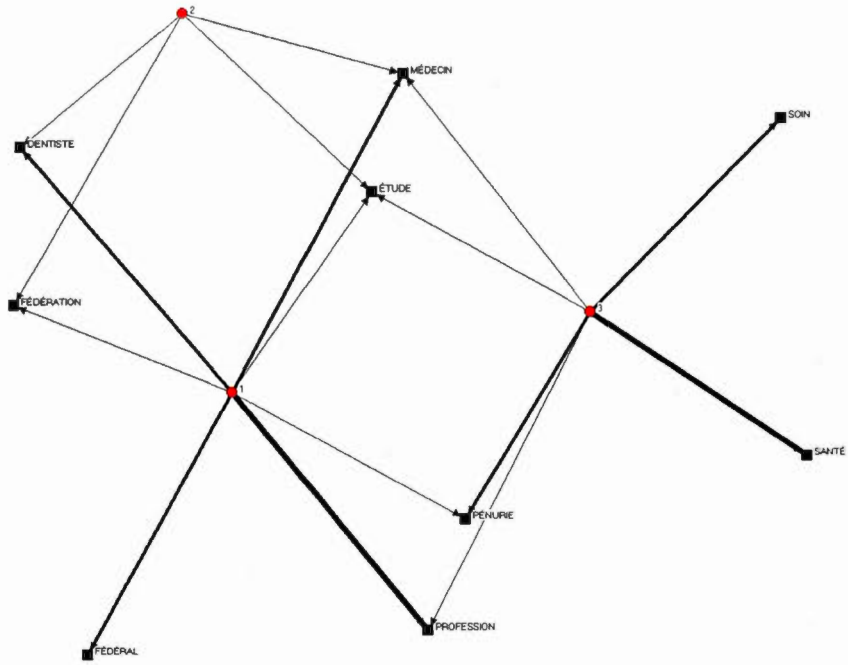
Représentation graphique du lexique de la classe 97.



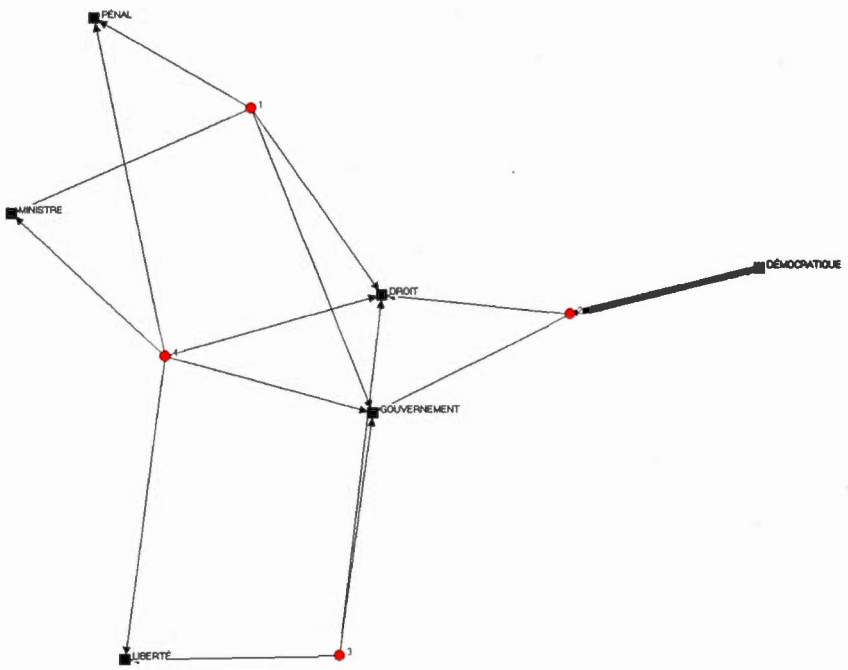
Représentation graphique du lexique de la classe 98.



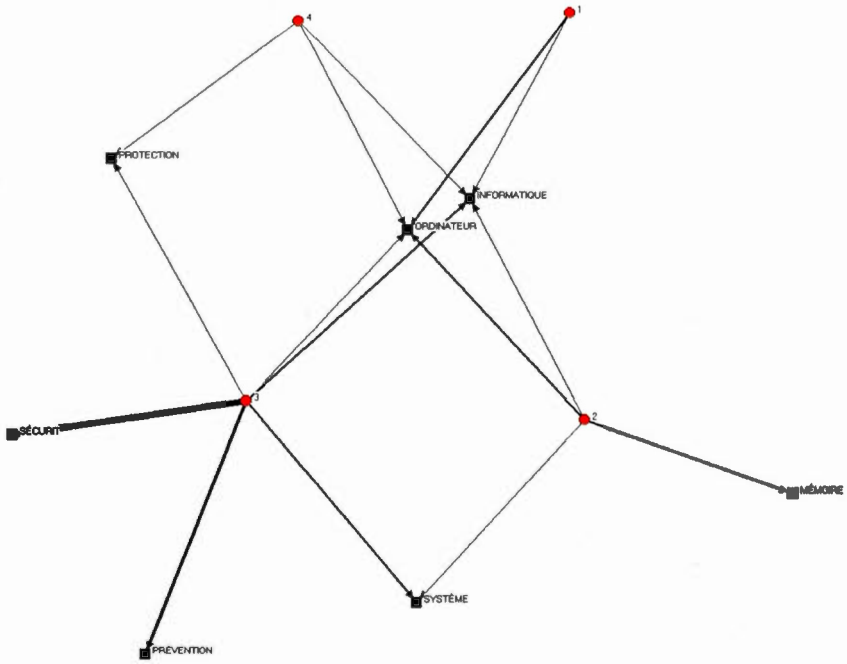
Représentation graphique du lexique de la classe 99.



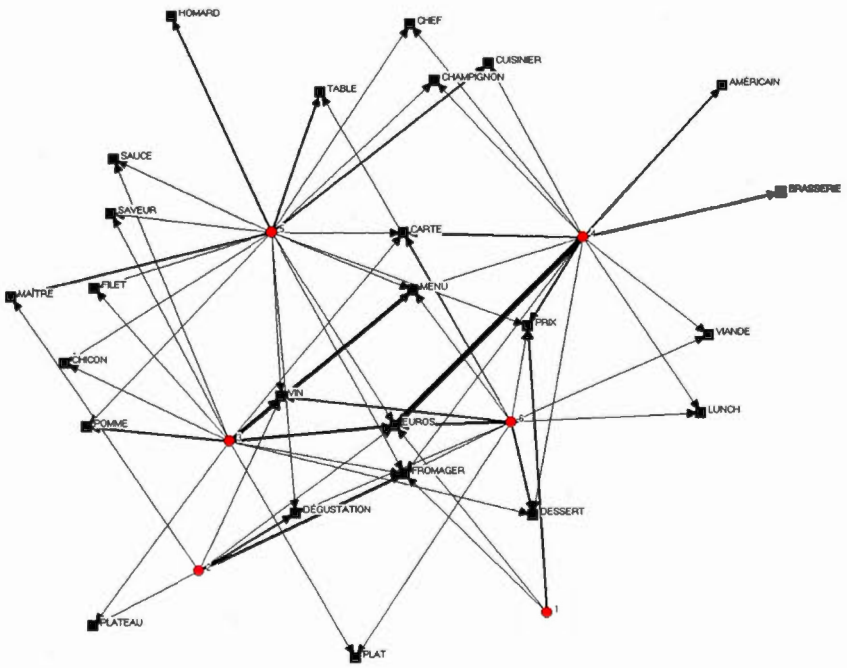
Représentation graphique du lexique de la classe 100.



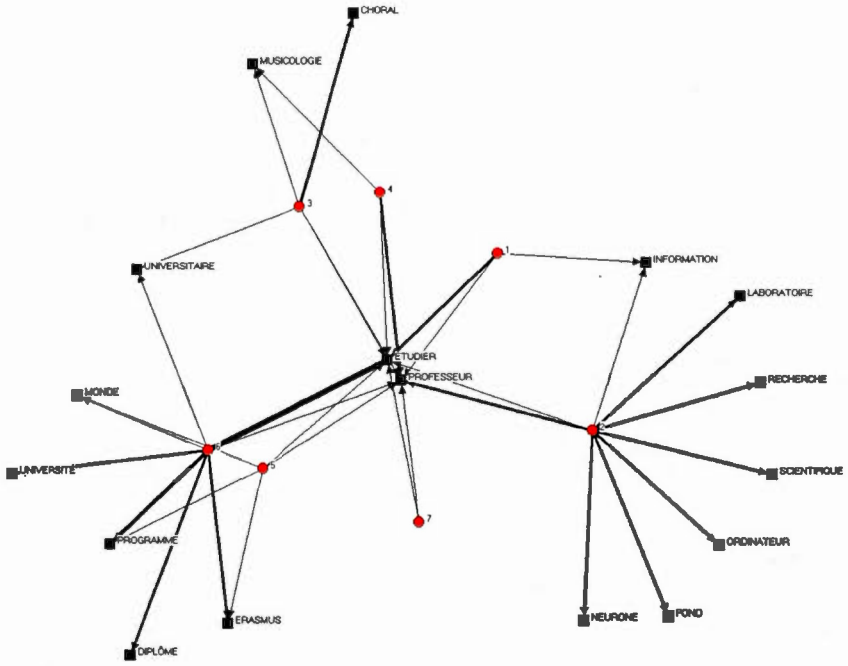
Représentation graphique du lexique de la classe 101.



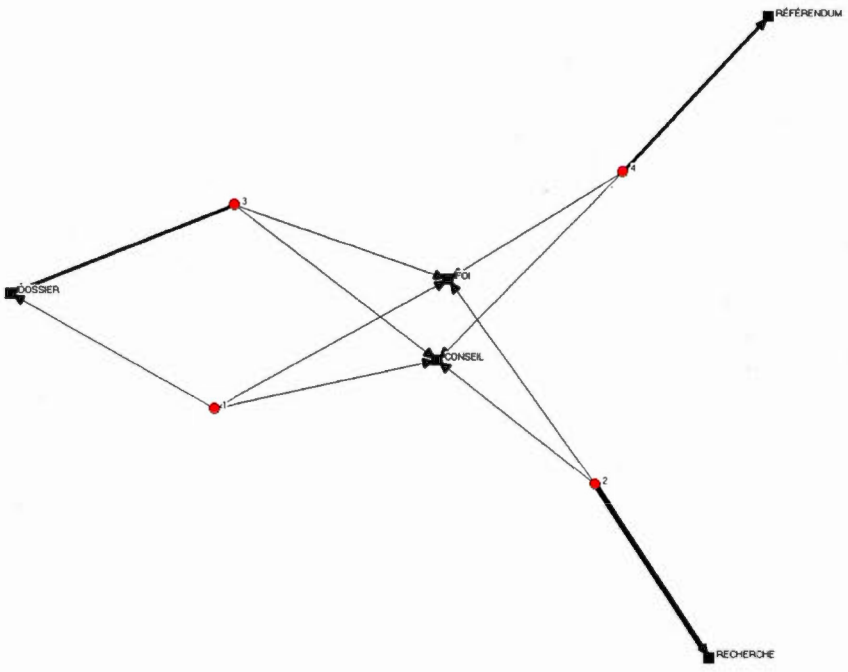
Représentation graphique du lexique de la classe 102.



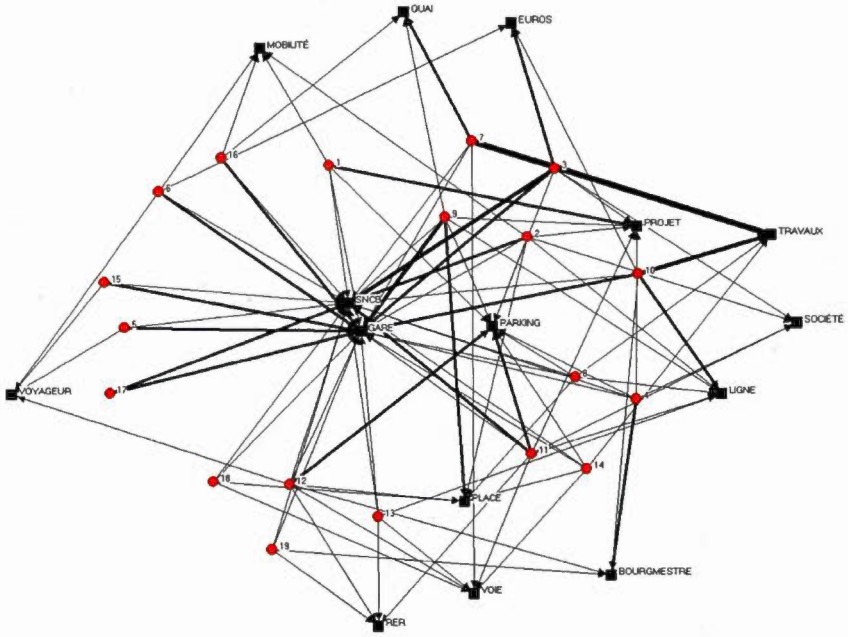
Représentation graphique du lexique de la classe 103.



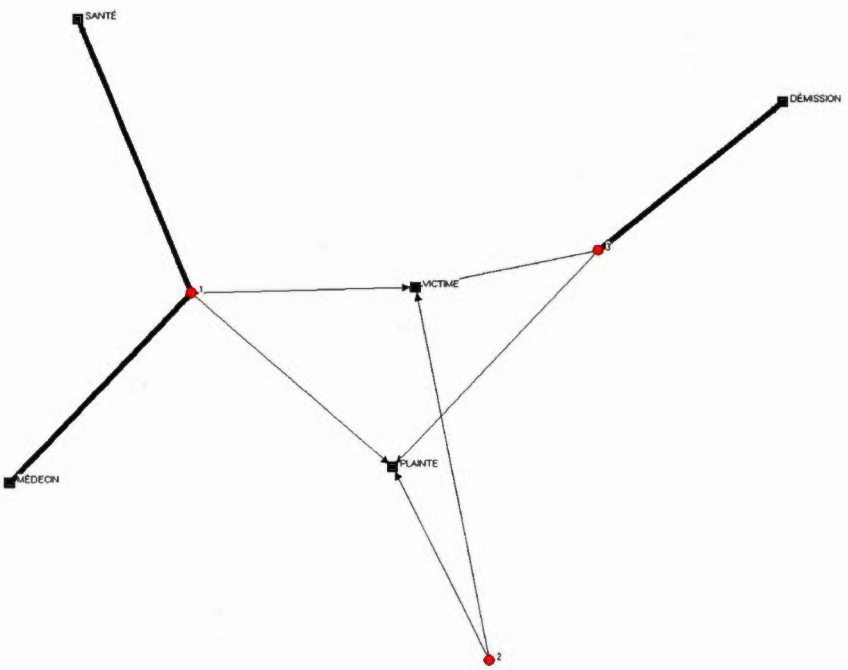
Représentation graphique du lexique de la classe 104.



Représentation graphique du lexique de la classe 105.

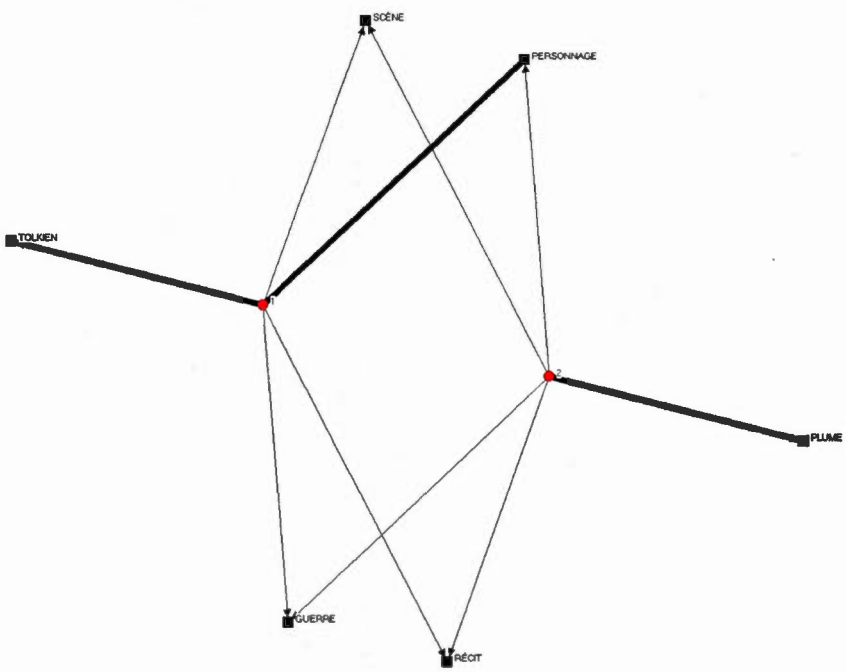


Représentation graphique du lexique de la classe 106.

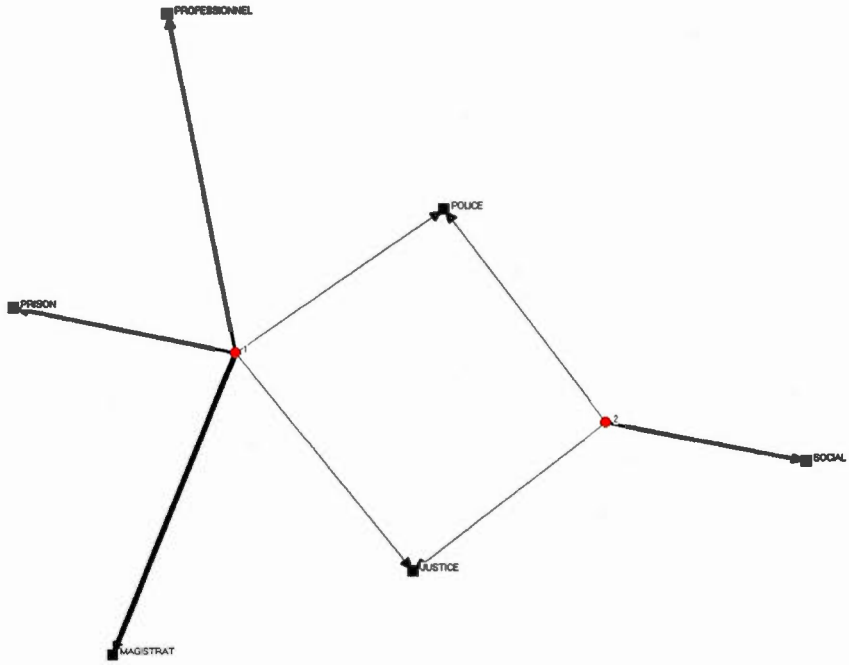


Représentation graphique du lexique de la classe 107.

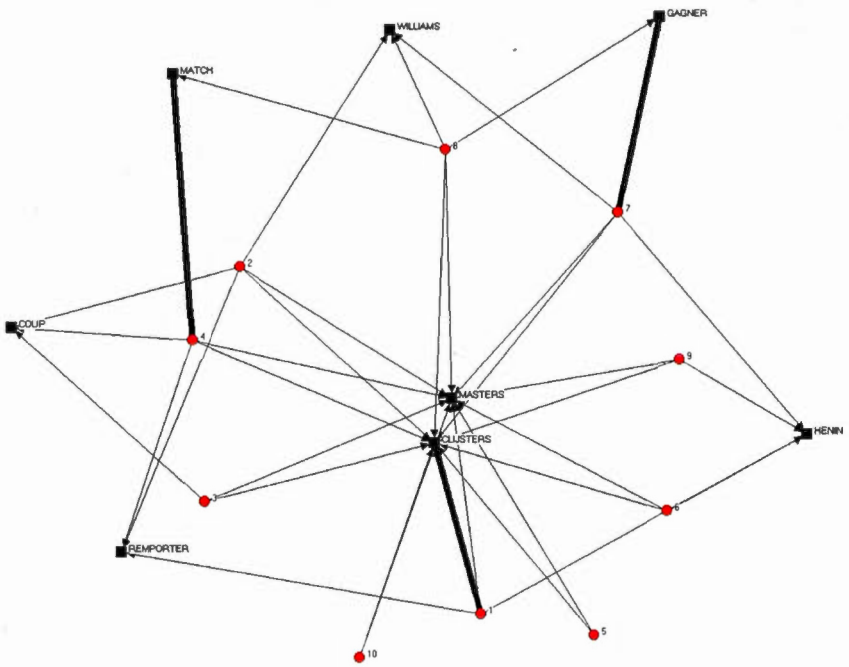
Représentation graphique du lexique de la classe 108.



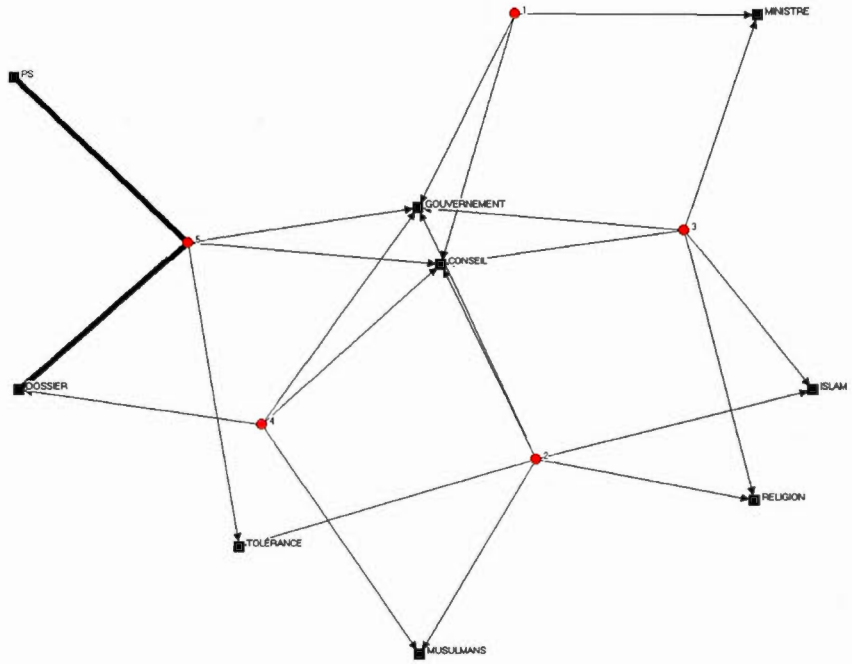
Représentation graphique du lexique de la classe 109.



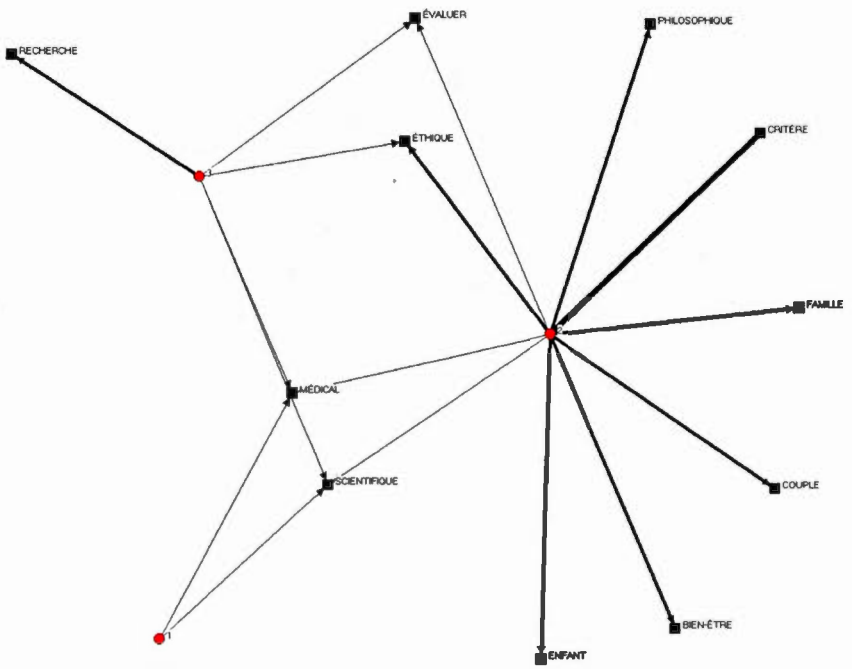
Représentation graphique du lexique de la classe 110.



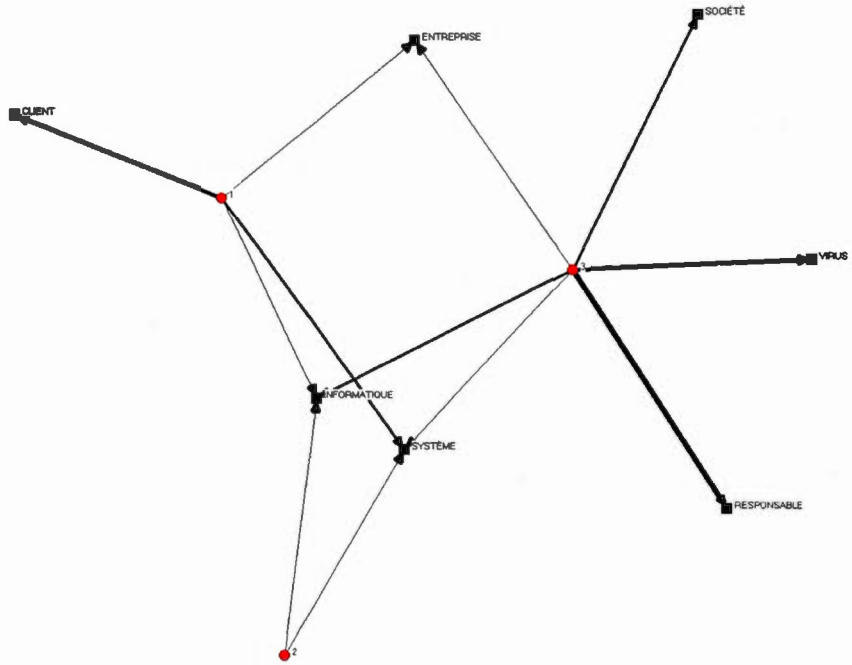
Représentation graphique du lexique de la classe 111.



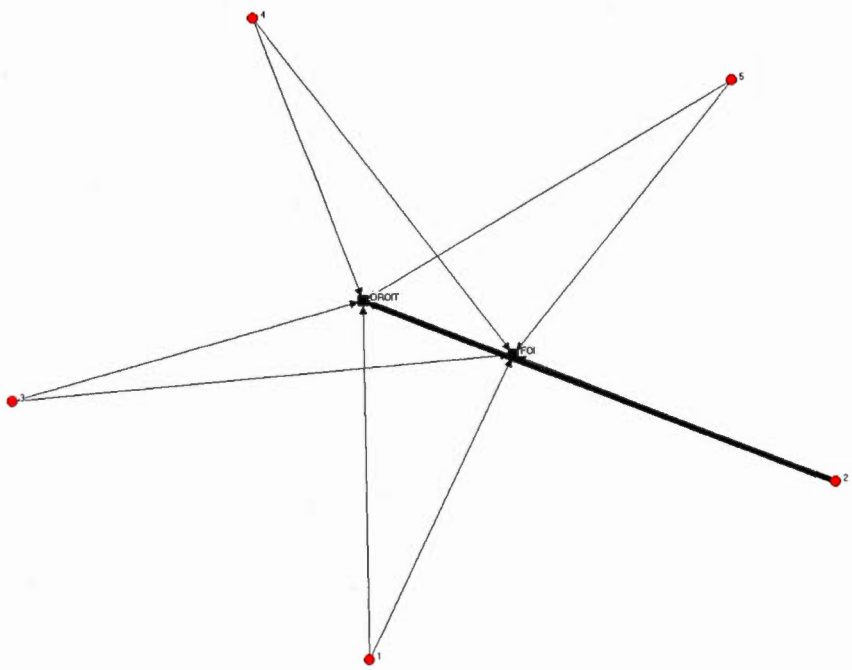
Représentation graphique du lexique de la classe 112.



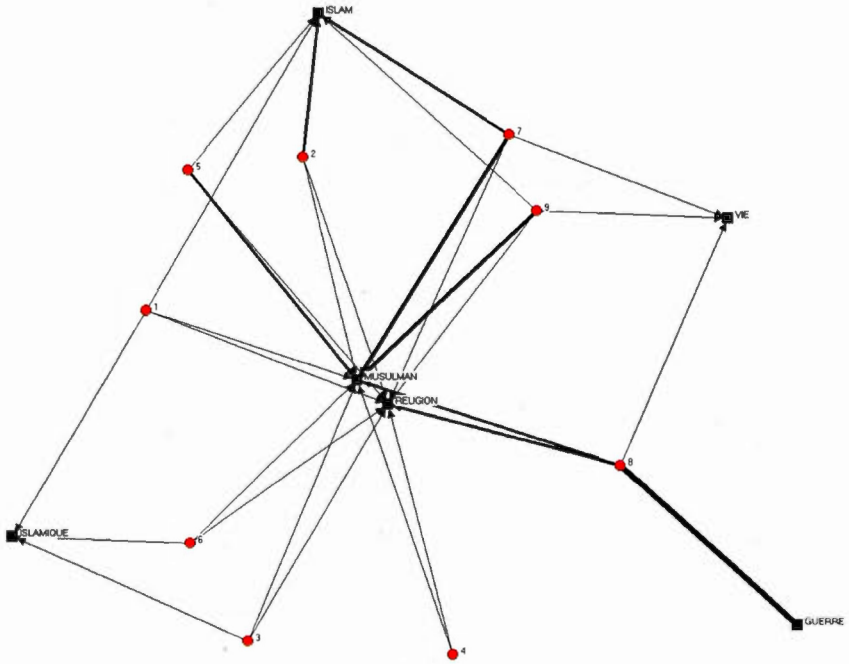
Représentation graphique du lexique de la classe 113.



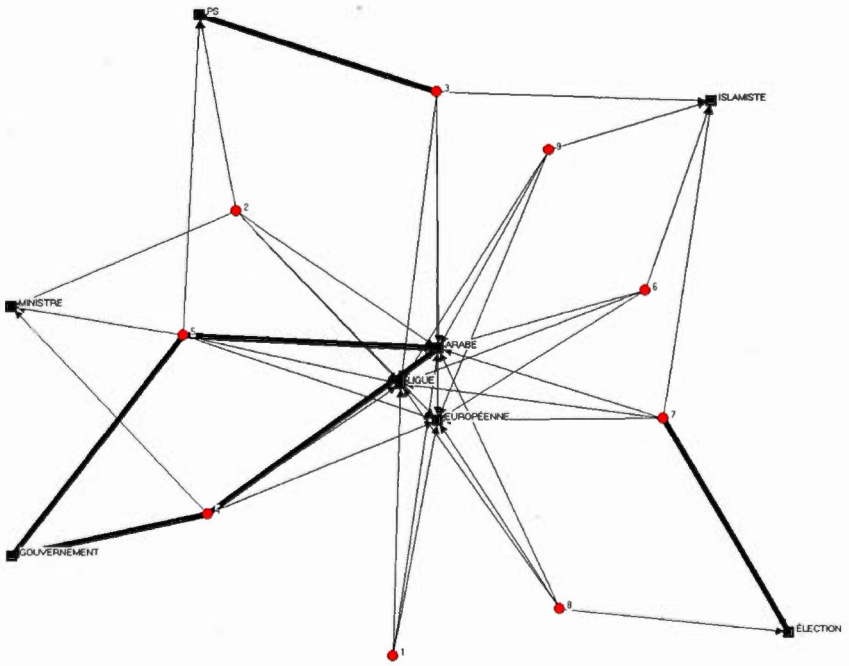
Représentation graphique du lexique de la classe 114.



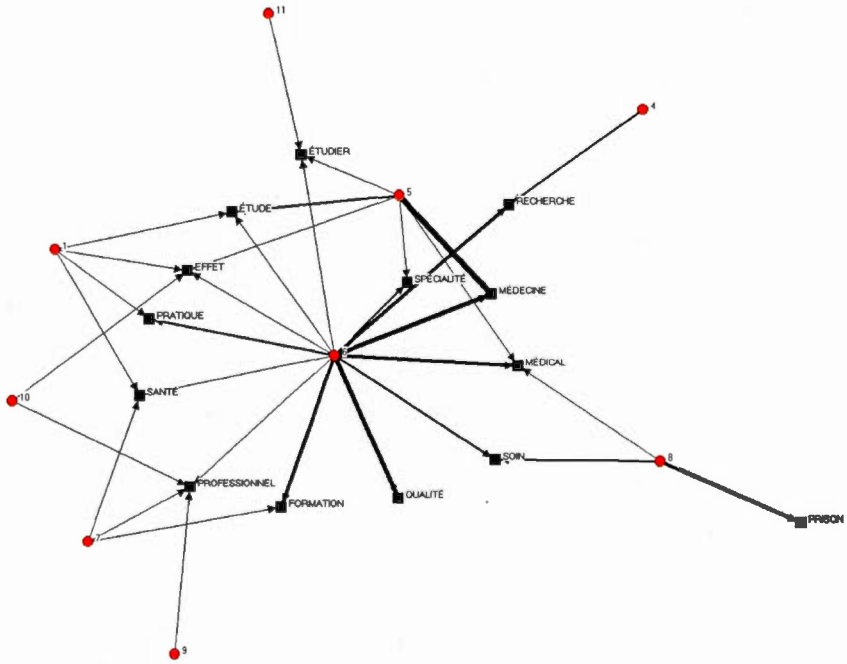
Représentation graphique du lexique de la classe 115.



Représentation graphique du lexique de la classe 116.



Représentation graphique du lexique de la classe 117.



Représentation graphique du lexique de la classe 118.

RÉFÉRENCES BIBLIOGRAPHIQUES

Aarnes, A. et Thompson, S. 1928. *The types of folk-tale: a classification and bibliography. Folklore fellows communications*, no 74. Helsinki : Suomalainen.

Aery, M., Ramamurthy, N. et Alp Aslandogan, Y. 2003. « Topic identification of textual data ». Technical report CSE-2003-25. Department of computer science and engineering, University of Texas at Arlington.

Aggarwal, C., Gates, S. et Yu, P. 2004. « On Using Partial Supervision for Text Categorization ». *IEEE Transactions on Knowledge and Data Engineering*. Vol. 16, no 2, pp. 245-255

Alexa, M. et Zuell, C. 1999a. *Commonalities, difference and limitations of text analysis software: the results of a review*. ZUMA arbeitsbericht, ZUMA: Mannheim.

Alexa, M. et Zuell, C. 1999b. *A review of software for text analysis*. ZUMA arbeitsbericht, ZUMA : Mannheim.

Alpaydin, E. 2004. *Introduction to machine learning*. Cambridge (Mass.) : MIT Press.

Archambeault, J. 2002. *Visualisation de l'évolution d'un domaine scientifique par l'analyse des résumés de publication à l'aide de réseaux neuronaux*. Mémoire de maîtrise, Montréal, École de Bibliothéconomie et des Sciences de l'Information, Université de Montréal.

Baeza-Yates, R. et Ribeiro, B. d. A. N. 1999. *Modern information retrieval*. New York : ACM Press / Addison-Wesley.

Balpe, J.P., A. Lelu et F. Papy. 1996. *Techniques avancées pour l'hypertexte.* Paris : Hermes.

Banerjee, S. et Pedersen, T. 2002. « An adapted lesk algorithm for word sense disambiguation using WordNet ». *In* Gelbukh, A. F. (dir. publ.). 2002. *Proceedings of the third international conference on computational linguistics and intelligent text processing.* Lecture notes in computer science, vol. 2276. London : Springer-Verlag, pp. 136-145.

Barry, C. A. 1998. « Choosing qualitative data analysis software: Atlas/ti and Nudist compared ». *Sociological Research Online*. Vol. 3, no 3. www.socresonline.org.uk/socresonline/3/3/4.html.

Basu, A., Watters, C. et Shepherd, M. 2003. « Support Vector Machines for Text Categorization ». *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03)*. Big Island, Hawaii.

Boeri, R. J. 2004. « Playing with taxinomies ». *EContent. Digital content strategies and ressources*. Vol. 27, no 12, p.12.

Borgatti, S.P., Everett, M.G. et Freeman, L.C. 2002. *Ucinet for Windows: software for social network analysis*. Harvard (Mass.) : Analytic Technologies.

Bremond, C. 1985. « Concept et thème ». *Poétique*. No 64, pp. 415-423.

Bremond, C., Landy, J. et Pavel, T. (dir. publ.). 1995. *Thematics. New approaches*. Albany : Suny Press.

Brunet, E. 2000. « Qui lemmatise dilemme attise ». *Lexicometrica*, no 2.

Brunet, E. 2002. *Le lemme comme on l'aime. In* Morin, A. et Sébillot, P (dir. publ.). *Actes des 6ièmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*. Saint-Malo, 13-15 mars 2002, vol. 1, pp. 221-232. Saint-Malo : IRISA/INRIA.

Callan, J.-P. 1994. « Passage retrieval evidence in document retrieval ». *In Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 302-310.

Cardoso-Cachopo, A. et Oliveira, A. L. 2003. « An empirical comparison of text categorization methods ». *In* Nascimento, M. A., Moura, E. S. et Oliveira, A. L. (dir. publ.). *String processing and information retrieval. Lecture notes in computer science, vol. 2857*. Berlin; New York : Springer-Verlag, pp. 183-196.

Chaar, S. 2003. « Extraction des segments thématiques pour la construction de résumé multi-document ». *RECITAL 2003*, 12 juin 2003, Batz-sur-Mer, France.

Chali, Y. 2005. « Topic detection of unrestricted texts: approaches and evaluations ». *Journal of Applied Artificial Intelligence*. Vol. 19, no 2, pp. 119-136.

Chuang, S.-L. et Chien, L.-F. 2003. « Enriching web taxonomies through subject categorization of query terms from search engine logs ». *Decision Support Systems*. Vol. 35, no 1, pp. 113-127

Clark, P. et Niblett, T. 1989. « The CN2 induction algorithm ». *Machine Learning Journal*. Vol. 3, no 4, pp. 261-283.

Davi, A., Haughton, D., Nasr, N., Shah, G. Skaletsky, M. et Spack, R. 2005. « A review of two text mining packages: SAS TextMining and WordStat ». *The American Statistician*. Vol. 59, no 1, pp. 89-103.

De Pasquale, J.-F. et Meunier, J.-G. 2003. « Categorisation techniques in computer-assisted reading and analysis of texts (CARAT) in the humanities ». *Computers and the Humanities*. Vol.37, no 1, pp.111-118.

Degenne, A. et Forsé, M. 1994. *Les Réseaux Sociaux.* Paris : Armand Colin.

Dejun, X. et Maosong, S. 2003. « A study on feature weighting in Chinese text categorization ». *In* Gelbukh, A. (dir. publ.). *Computational linguistics and intelligent text processing. Lecture notes in computer science, vol. 2588*. Berlin; New York : Springer-Verlag, pp. 592-601.

Despres, C. et Chauvel, D. (dir. publ.). 2000. *Knowledge horizons. The present and the promise of knowledge management*. Boston (Mass.) : Butterworth Heinemann

Diederich, J., Kindermann, J., Leopold, E., et Paass, G. 2003. « Authorship attribution with support vector machines ». *Applied Intelligence*. Vol. 19, no 1-2, pp. 109-123.

Ding, Y. et Foo, S. 2002a. « Ontology research and development. Part 1 - a review of ontology generation ». *Journal of Information Science*. Vol. 28, no 2, pp. 123-136.

Ding, Y. et Foo, S. 2002b. « Ontology research and development. Part 2 - a review of ontology mapping and evolving ». *Journal of Information Science*. Vol. 28, no 5, pp. 375-388.

Dittenbach, M., Berger, H. et Merkl, D. 2004. « *Improving domain ontologies by mining semantics from text* ». *In* Hartmann, S. et Roddick, J. (dir. publ.). *Proceedings of the 1st Asia-Pacific Conference on Conceptual Modelling (APCCM 2004)*. Australian Computer Society, vol. 31, pp. 91-100.

Feldman, S. 2004. « Why categorize? ». *KMWorld. Content document and knowledge management*. Vol. 13, no 9, pp. 8-10.

Fellbaum, C. (dir. publ.). 1998. *Wordnet: an electronic lexical database*. Cambridge (Mass.) : MIT Press.

Forest, D. 2002. *Lecture et analyse de textes philosophiques assistées par ordinateur : application d'une approche classificatoire mathématique à l'analyse thématique du* Discours de la méthode *et des* Méditations métaphysiques *de Descartes*. Mémoire de maîtrise, Montréal, Université du Québec à Montréal.

Forest, D. 2004. *Automated text categorization: theory and application to computer-assisted text analysis in the humanities*. The Third Canadian Symposium on Text Analysis Research (CaSTA), symposium organisé par le groupe de recherche *TAPoR* (*Text Analysis Portal for Research*), 19-21 novembre 2004, Hamilton, Ontario.

Forest, D. 2005. *Application de techniques de forage de textes à des fins de gestion et d'analyse thématique de documents textuels non structurés*. Département d'informatique, Université du Québec à Montréal, 7 juin 2005.

Forest, D. 2005. *La classification et la catégorisation automatiques : application à l'analyse et à la gestion automatisées des documents textuels. Congrès annuel de l'Association Canadienne des Sciences de l'Information (ACSI)*, 2-5 juin 2005, London, Ontario.

Forest, D. et Meunier, J.-G. 2000. *La classification mathématique des textes : un outil d'assistance à la lecture et à l'analyse de textes philosophiques. In* Rajman, M. & Chappelier, J.-C. (eds.). *Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles*, 9-11 mars 2000, EPFL, Lausanne, Suisse. Volume 1, pp. 325-329.

Forest, D. et Meunier, J.-G. 2004. *Classification et catégorisation automatiques : application à l'analyse thématique des données textuelles. In* Purnelle, G., Fairon, C. et Dister, A. (Dir. publ.). *Le poids des mots. Actes des 7ièmes Journées internationales d'Analyse statistique des Données Textuelles.* Louvain-la-Neuve : Presses Universitaires de l'Université Catholique de Louvain. Volume 1, pp. 434-444.

Frakes, W. B. et Baeza-Yates, R. (dir. publ.). 1992. *Information retrieval : data structures and algorithms*. Englewood Cliffs : Prentice-Hall.

Garcia, E. C. 1975. *The role of theory in linguistic analysis : the spanish pronoun system.* New York : Elsevier.

Gelbukh, A. (dir. publ.). 2003. *Computational linguistics and intelligent text processing. Lecture notes in computer science, vol. 2588*. Berlin; New York : Springer-Verlag.

Giora, R. 1985. « Notes toward a theory of text coherence ». *Poetics Today*. Vol 6, no, 4.

Grabmeier, J. et Rudolph, A. 2002. « Techniques of cluster algorithms in data mining ». *Data mining and knowledge discovery*. Vol. 6, pp. 303-360.

Grossberg, S. et Carpenter, G. A. 1987. « A massively parallel architecture for a self-organizing neural pattern recognition machine ». *Computer Vision, Graphics, and Image Processing*. No 37, pp. 54-115.

Grossberg, S., Carpenter, G.A. et Rosen, D.B. 1991. « ART2-A : An adaptive resonance algorithm for rapid category learning and recognition ». *Neural Networks*. No 4, pp. 493-504.

Grossberg, S., G.A. Carpenter, N. Markuzon et J.H. Reynolds. 1991. « Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system ». *Neural Networks*. No 4, pp. 759-771.

Han-Joon, K. et Lee, S. G. 2003. « Building topic hierarchy based on fuzzy relations ». *Neurocomputing*. Vol. 51, pp. 481-486.

Haruechaiyasak, C. *et al.* 2002. « Web document classification based on fuzzy association ». *Proceedings of the 26th IEEE computer society international computer software and applications conference (COMPSAC)*. pp. 487-492.

Havre, S., Hetzler, E., Whitney, P. et Nowell, L. 2002. « ThemeRiver : visualizing thematic changes in large document collections ». *IEEE Transactions on Visualization and Computer Graphics*. Vol. 8, no 1, pp. 9-20.

He, J., Tan, A.-H. et Tan, C.-L. 2003. « On machine learning methods for Chinese document categorization ». *Applied Intelligence*. Vol. 18, pp. 311-322.

Hearst, M. 1999. « Untangling text data mining ». *In Proceeding of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*. University of Maryland, 20-26 juin.

Hearst, M. A. et Plaunt, C. 1993. « Subtopic structuring for full-length document access ». *Proceedings of the 16th annual international ACM/SIGIR conference on research and development in information retrieval*, pp. 59-68.

Hélie, S., Proulx, R., et Lefebvre, B. 2006. « JPEX: A psychologically plausible Joint Probability Extractor ». *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*. (soumis)

Hirst G. et St-Onge D. 1998. *Lexical chains as representations of context for the detection and correction of malapropisms. In* Fellbaum, C. (dir. publ.). *Wordnet: an electronic lexical database*. Cambridge (Mass.) : MIT Press, pp. 305-332.

Hogenraad, R. 2002. *Moving targets: The making and modeling of a theme. In* Louwerse, M. et Van Peer, W. (dir. publ.). 2002. *Thematics: Interdisciplinary Studies*. Pays-Bas : John Benjamins Publishing Company, pp. 353-376.

Hovy, E. et Radev, D. (dir. publ.). 1998. *Intelligent text summarization: papers from the AAAI spring symposium (technical reports, vol. SS-98-06)*. Menlo Park (Calif.) : AAAI Press.

Hule, D. A, Pederson, J. O. et Schütze, H. 1996. « Method combination for document filtering ». In Frei, H.P., Harman, D. et Schaüble, P. et Wilkinson, R. (dir. publ.). *Proceedings of SIGIR-96, 19th ACM international conference on research and development in information retrieval*. New York : ACM Press, pp. 279-288.

Ishizuka, M. et Satter, A. (dir. publ.). 2002. *PRICAI 2002: trends in artificial intelligence. Lecture notes in artificial intelligence, Vol. 2417*. Berlin; New York : Springer-Verlag.

Jackson, P. et Moulinier, I. 2002. *Natural language processing for online applications: text retrieval, extraction, and categorization*. Amsterdam : John Benjamins Publishing Company.

Jain, A. K., Murty, M. N. et Flynn, P. J. 1999. « Data clustering: a review ». *ACM Computing Surveys*. Vol. 31, no 3, pp. 264-323.

Jiang, J. J. et Conrath, D. W. 1997. « Semantic similarity based on corpus statistics and lexical taxonomy ». *In Proceedings of the 10th international conference on research in computational linguistics (ROCLING)*. Tapei, Taiwan.

Joachims, T. 2002. *Learning to classify text using support vector machines*. Dordrecht: Kluwer Academic Publishers.

Kastberg Sjöblom, M. et Brunet, E. 2000. *La thématique. Essai de repérage automatique dans l'œuvre d'un écrivain. In* Rajman, M. et J.-C. Chappelier (dir. publ.). *Actes des 5e Journées internationales d'Analyse statistique des Données Textuelles (JADT)*. Lausanne, 9-11 mars 2000, vol. 2, pp. 457-466. Lausanne : EPFL.

KaszKiel, M. et Zobel, J. 2001. « Effective ranking with arbitrary passages ». *Journal of the American Society for Information Science*. Vol 52, no 4, pp. 344-364.

Kelle, U. 1997. « Theory building in qualitative research and computer programs for the management of textual data ». *Sociological Research Online*. Vol. 2, no 2.

Kim, J. et Kim, M.-H. 2004. « An Evaluation of Passage-Based Text Categorization ». *Journal of Intelligent Information Systems*. Vol. 23, no 1, pp. 47-65.

Kintsch, W. 2002. *On the notions of theme and topic in psychological proess models of text comprehension. In* Louwerse, M. et Van Peer, W. (dir. publ.). *Thematics: Interdisciplinary Studies*. Amsterdam : John Benjamins Publishing Company, pp. 157-170.

Kintsch, W. et Van Dijk, T. A. 1978. « Toward a model of text comprehension and production ». *Psychological Review*, vol. 85, no 5, pp. 363-394.

Kohonen, T. 2001. *Self-Organizing Maps*. Berlin : Springer.

Krause, J. 1996. « Principles of content analysis for information retrieval systems ». *In* C. Züll, Harkness, J. et Hoffmeyer-Zlotnik, J.H.P. (dir. publ.). *Text analysis and computers*. Mannheim : ZUMA (ZUMA Nachrichten Spezial), pp. 77-100.

Krishnapuram, R. et Kummamuru, K. 2003. « Automatic taxonomy generation: issues and possibilities ». *Proceedings of fuzzy sets and systems (IFSA), LNCS 2715*. Heidelberg : Springer-Verlag, pp. 52-63.

Lallich-Boidin, G. et Maret, D. 2005. *Recherche d'information et traitement de la langue.* Lyon : Presses de l'ENSSIB.

Leacock C. et Chodorow M. 1998. *Combining local context and WordNet similarity for word sense identification. In* Fellbaum, C. (dir. publ.). *Wordnet: an electronic lexical database.* Cambridge (Mass.) : MIT Press, pp. 265-283.

Lebart, L. et A. Salem. 1988. *Analyse statistique des données textuelles.* Paris : Dunod.

Lebart, L., Salem, A. et Berry, L. 1998. *Exploring Textual Data.* Dordrecht: Kluwer Academic Publishers.

Lee, H., Kay, J., Kang, B. H. et Rosebrock, U. 2002. « A comparative study on statistical machine learning algorithms and thresholding strategiesfor automatic text categorization ». *In* Ishizuka, M. et Satter, A. (dir. publ.). *PRICAI 2002: trends in artificial intelligence. Lecture notes in artificial intelligence, Vol. 2417.* Berlin; New York : Springer-Verlag, pp. 444-453.

Lewis, D. D. 1995. « Evaluating and optimizing autonomous text classification systems ». *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval.* Seattle, Washington, pp. 246-253.

Lewis, D. D. et Ringuette, M. 1994. « A comparison of two learning algorithms for text categorization ». *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 81-93.

Lin, D. 1998. « An information-theoretic definition of similarity ». *In* Shavlik, J. W. (dir. publ.). *Proceedings of the fifteenth international conference on machine learning.* San Francisco (CA) : Morgan Kaufmann Publishers, pp. 296-304.

Louwerse, M. et Van Peer, W. (dir. publ.). 2002. *Thematics: Interdisciplinary Studies.* Amsterdam : John Benjamins Publishing Company

Luhn, H. P. 1957. « A statistical approach to mechanized encoding and searching of literary information ». *IBM Journal of Research and Development.* Vol 1, no 4, pp. 309-317.

Mani, I. 2001. *Automatic summarization.* Amsterdam : John Benjamins Publishing Company.

Mani, I. et Maybury, M. T. (dir. publ.). 1999. *Advances in automatic text summarization.* Cambridge (Mass.) : MIT Press.

Manning, C. D. et H. Schütze. 1999. *Foundations of statistical natural language processing.* Cambridge (Mass.): MIT Press.

Maron, M. 1961. « Automatic indexing: an experimental inquiry ». *Journal of the Association for Computing Machinery.* Vol. 8, no 3, pp. 404-417.

Martin, É. 1995. *Thème d'étude, étude de thème. In* Rastier, F. (dir. publ.). *L'analyse thématique des données textuelles : l'exemple des sentiments*. Paris : Didier érudition, pp. 13-24.

Massey, L. 2003a. « Evaluating quality of text clustering with ART1 ». *Proceedings of 2003 international joint conference on neural networks*. Vol. 2, pp. 1402-1407.

Massey, L. 2003b. « On the quality of ART1 text clustering ». *Neural Networks*. Vol. 16, no. 5-6, pp. 771-778.

Memmi, D. 2000. *Le modèle vectoriel pour le traitement de documents*. Grenoble : Cahiers Leibniz, no 2000-14.

Meunier, J.-G. 1995. *La Lecture et l'analyse de texte assistée par ordinateur : La chaîne d'analyse. Cahiers de recherche du Laboratoire d'ANnalyse Cognitive de l'Information*. Vol. 6.

Meunier, J.-G., Forest, D. et Biskri, I. 2005. *Classification and categorization in computer assisted reading and analysis of texts. In* Lefebvre, C. et Cohen, H. (dir. publ.). 2005. *Handbook of categorization in cognitive science*. New York: Elsevier, pp. 955-978.

Meunier, J.-G., Remaki, L. et Forest, D. 1999. « Use of classifiers in Computer-Assisted Reading and Analysis of Text (CARAT) ». *Actes du colloque international CISST 1999 (The 1999 International Conference on Imaging Science, Systems and Technology)*, 28 juin-1er juillet 1999, Las Vegas, U.S.A.

Miller, G.A. *et al.* 1993. *Five papers on WordNet*. Technical report, Princeton University.

Miller, N. E., Chung Wong, P., Brewster, M. et Foote, H. 1998. « Topic-Islands - A wavelet-based text visualization system ». *IEEE Visualization - Proceedings of the conference on Visualization '98*, pp. 189-196.

Missikoff, M., Velardi, P. et Fabriani, P. 2003. « Text mining techniques to automatically enrich a domain ontology ». *Applied Intelligence*. Vol. 18, pp. 323-340.

Moffat, A., Sacks-Davis, R., Wilkinson, R. et Zobel, J. 1994. « Retrieval of partial documents ». In *NIST Special Publication 500-215: The Second Text REtrieval Conference (TREC 2)*, pp. 181-190.

Muller, C. 1968. *Initiation à la statistique linguistique*. Paris : Larousse.

Nascimento, M. A., Moura, E. S. et Oliveira, A. L. (dir. publ.). 2003. *String processing and information retrieval. Lecture notes in computer science, vol. 2857*. Berlin; New York : Springer-Verlag.

Nauck, U. 1999. *Design and Implementation of a Neuro-Fuzzy Data Analysis Tool in Java.* Thèse de doctorat, Braunschweig, Technische Universität Braunschweig.

Nault G., V. Rialle et J.-G. Meunier. 1999. « PROGEN : a genetic-based semi-automatic hypertext construction tool - first steps and experiment ». *In* Eiben, A. E., M. H. Garzon, V. Honavar, M. Jakiela et R. E. Smith (dir. publ.). *GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference, July 13-17, 1999, Orlando, Florida USA*. San Francisco : Morgan Kaufman Publishers.

Nault, G. 2000. *Approche cognitive de l'hypertextualisation semi-automatique. Effets sur la conception d'un système d'assistance interactive fondé sur un optimiseur émergentiste.* Thèse de doctorat, Montréal, Université du Québec à Montréal.

Patwardhan, S. 2003. *Incorporating dictionary and corpus information into a context vector measure of semantic relatedness*. Master of science thesis, department of computer science, University of Minnesota, Duluth.

Popping, R. 2000. *Computer-assisted text analysis*. London : Sage.

Porter, M. F. 1980. « An algorithm for suffix stripping ». *Program*. Vol. 14, no 3, pp. 130-137.

Pottier, B. 1974. *Linguistique générale : théorie et description*. Paris : Klincksieck.

Prince, G. 1985. « Thématiser ». *Poétique*. No 64, pp. 425-433.

Propp, V. 1928/1968. *Morphology of the folktale*. Austin : Texas University Press.

Quinlan, J. R. 1993. *C4.5: Programs for Machine learning*. San Mateo (Calif.) : Morgan Kaufman.

Rastier, F. 1987/1996. *Sémantique interprétative*. Paris : Presses Universitaires de France.

Rastier, F. 1995. (dir. publ.). *L'analyse thématique des données textuelles : l'exemple des sentiments*. Paris : Didier érudition.

Rastier, F. 2001. *Arts et sciences du texte*. Paris : Presses Universitaires de France.

Rastier, F. *et al.* (dir. publ.). 1995. *L'analyse thématique des données textuelles : l'exemple des sentiments*. Paris : Didier Érudition.

Rastier, F. *et al.* 1994. *Sémantique pour l'analyse. De la linguistique à l'informatique*. Paris : Masson.

Reinhart, T. 1981. « Pragmatics and linguistics: an analysis of sentence topics ». *Philosophica*. No 27, pp. 53-93.

Resnik P. 1999. « Semantic similarity in a taxonomy: an information-based measure and its applications to problems of ambiguity in natural language ». *Journal of Artificial Intelligence Research.* Vol. 11, pp. 95-130.

Riloff, E. 1995. « Little words can make a big difference for text classification ». *Proceedings of the 18th International Conference on Research and Development in Information Retrieval.* New-York : ACM Press, pp. 130-136.

Rimmon-Kenan, S. 1985. « Qu'est-ce qu'un thème? ». *Poétique.* No 64, pp. 397-406.

Rivero, L. C., Doorn, J. H. et Ferraggine, V. E. (dir. publ.). 2005. *The encyclopedia of database technologies and applications.* Hershey : Idea Group Publishing.

Robert, A. D. et Bouillaguet, A. 1997. *L'analyse de contenu.* Paris : Presses Universitaires de France.

Rosenblatt, F. 1958. « The perceptron: a probabilistic model for information storage and organisation in the brain ». *Psychological Review.* No 65, pp. 386-408.

Rossignol, M. 2005. *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle.* Thèse de doctorat, Rennes, Université de Rennes 1.

Rossignol, M. et Sébillot, P. 2002. *Automatic generation of sets of keywords for theme characterization and detection. In* Morin, A. et Sébillot, P (dir. publ.). *Actes des 6ièmes Journées internationales d'Analyse statistique des Données Textuelles (JADT).* Saint-Malo, 13-15 mars 2002, vol. 2, pp. 653-664. Saint-Malo : IRISA/INRIA.

Rossignol, M. et Sébillot, P. 2003. « Extraction statistique sur corpus de classes de mots-clés thématiques ». *TAL (Traitement Automatique des Langues).* Vol. 44, no 3, pp. 217-246.

Rossignol, M. et Sébillot, P. 2005. « Combining statistical data analysis techniques to txtract Topical keyword classes from corpora ». *IDA (Intelligent Data Analysis).* Vol. 9, no 1, pp. 105-127.

Roy, T et Beust, P. 2004. *ProxiDocs : un outil de cartographie et de catégorisation thématique de corpus. In* Purnelle, G., Fairon, C., et Dister, A. (dir. publ.). *Le poids des mots : Actes des 7e Journées internationales d'Analyses statistique des Données Textuelles (JADT).* 10-12 mars 2004, Louvain-la-Neuve : Presses Universitaires de Louvain, Volume 2, pp. 978 à 986.

Ruggles, R. 1997. *Knowledge tools: using technology to manage knowledge better.* http://www.businessinnovation.ey.com/mko/html/toolsrr.html

Ruiz, E. et Srinivasan, P. 1998. *Automatic text categorization using Neural Networks*. In Efthimiadis, E. (Dir. publ.). *Advances in classification research, vol. 8 : Proceedings of the 8th ASIS SIG/CR classification research workshop*. New Jersey : Information Today, pp. 59-72.

Salton, G. 1989. *Automatic Text Processing*. Reading (Mass.) : Addison-Wesley.

Salton, G. et McGill, M. 1983. *Introduction to Modern Information Retrieval*. New-York: McGraw-Hill.

Salton, G., J. Allan et C. Buckley. 1994. « Automatic structuring and retrieval of large text File ». *Communications of the ACM*. Vol. 37, no 2, pp. 97-107.

Schachter, P. 1973. « Focus and relativization ». *Language*. No 18, pp. 19-46.

Schultz, C. K. 1968. *H.P. Luhn: Pioneer of information science - Selected Works*. London : Macmillan.

Scott, J. 1991. *Social network analysis: a handbook*. Newbury Park (CA) : Sage Publications.

Sebastiani, F. 1999. « A tutorial on automated text categorization ». *In* Amandi, A. et Zunino, A. (dir. publ.). *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence*, pp. 7-35.

Sebastiani, F. 2002. « Machine learning in automated text categorization ». *ACM Computing Surveys*. Vol. 34, no 1, pp. 1-47.

Sebastiani, F. 2005a. *Text categorization. In* Rivero, L. C., Doorn, J. H. et Ferraggine, V. E. (dir. publ.). *The encyclopedia of database technologies and applications*. Hershey : Idea Group Publishing.

Sebastiani, F. 2005b. *Text categorization. In* Zanasi, A. (dir. publ.). *Text mining and its applications*. Southampton : WIT Press.

Shamsfard, M. et Barforoush, A. A. 2004. « Learning ontologies from natural language texts ». *International Journal of Human-Computer Studies*. Vol. 60, no 1, pp. 17 - 63.

Sollors, W. 1993. *The return of thematic criticism*. Cambridge ; Harvard University Press.

Spangler, S. et Kreulen, J. 2002. « Interactive methods for taxonomy editing and validation ». *Proceedings of the 11th international conference on information and knowledge management*. New York : ACM Press, pp. 665-668.

Spinoza. 1955 (1670). *Traité des autorités théologique et politique. In Œuvres complètes*. Paris : Bibliothèque de la pléiade.

Staab, S. and R. Studer. 2004. *Handbook on ontologies*. Berlin; New York : Springer-Verlag.

Stamatatos, E., Kokkinakis, G., et Fakotakis, N. 2000. « Automatic text categorization in terms of genre and author ». *Computationnal Linguistics*. Vol. 26, no 4, pp. 471-495.

TDT. 2004. *TDT 2004 : annotation manual*. Version 1.2. www.ldc.upenn.edu/Projects/TDT2004.

Thomashevsky, B. 1925. *Thematics*. *In* Lemon, L. T. et Reis, M. J. (dir. publ.). 1965. *Russian formalist criticism*. Lincoln : University of Nebraska Press, pp. 61-98.

Thompson, S. 1946. *The folktale*. Berkeley : University of California Press.

Torres-Moreno, J.-M., Velazquez-Morales, P. et Meunier, J.-G. 2002. *Condensés de textes par des méthodes numériques*. *In* Morin, A. et Sébillot, P. (dir. publ.). *Actes des 6ièmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*. Saint-Malo, 13-15 mars 2002, vol. 2, pp. 723-734. Saint-Malo : IRISA/INRIA.

Touzet, C. 1992. *Une introduction aux réseaux de neurones artificiels*. Paris : EC2 Éditeur.

Van Dijk, T. A. 1972. *Some aspects of text grammars. A study in theoretical linguistics and poetics*. The Hague : Mouton.

Van Dijk, T. A. et Kintsch, W. 1983. *Strategies of discourse comprehension*. New York : Academic Press.

Van Rijsbergen, C. J. 1979. *Information retrieval*. London: Butterworths.

Veronis, J., N. Ide et S. Harié. 1990. « Utilisation de grands réseaux de neurones comme modèles de représentation des relations sémantiques ». *Actes des 3èmes Journées Internationales « Les Réseaux Neuro-Mimétiques et leurs Applications »*, pp. 527-538.

Voorhees, E. M. et Harman, D. (dir. publ.). 2005. *TREC : experiment and evaluation in information retrieval*. Cambridge (Mass.) : MIT Press.

Warner, A. J. 2002. « Metadata and taxonomies for a more flexible information architecture ». *3rd Annual Information Architecture Summit (ASIST)*. Baltimore, Maryland, 16 mars 2002.

Wasserman, S., et Faust, K. 1994. *Social network analysis: methods and applications*. New York : Cambridge University Press.

Weigel, F. 2004. *What is knowledge management?* http://www.gdrc.org/kmgmt/what-is-km.html

Weiss, S. M., Indurkhya, N., Zhang, T. et Damereau, F. J. 2005. *Text mining. Predictive methods for analyzing unstructured information*. Berlin; New York : Springer-Verlag.

Wu, Z. et Palmer, M. 1994. « Verbs semantics and lexical selection ». *In Proceedings of the 32nd annual meeting on Association For Computational Linguistics*. Las Cruces, New Mexico, pp. 133-138.

Yang, Y. 1999. « An evaluation of statistical approaches to text categorization ». *Information Retrieval*. Vol. 1, no 1-2, p. 69-90.

Yang, Y. et Liu, X. 1999. « A re-examination of text categorization methods ». *Proceedings of SIGIR-99, the 22nd ACM international conference on research and development in information retrieval*. New York : ACM Press, pp. 42-49.

Yang, Y. et Pederson, J. O. 1997. « A comparative study on feature selection in text categorization ». *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412-420.

Zadeh, L. A. 1965. « Fuzzy-Sets ». *Information and Control*. No 8, pp. 338-353.

Zanasi, A. (dir. publ.). 2005. *Text mining and its applications*. Southampton : WIT Press.

Zeimpekis, D. et Gallopoulos, E. 2005. *TMG: A MATLAB toolbox for generating term-document matrices from text collections*. Technical Report HPCLAB-SCG 1/01-05, Computer Engineering & Informatics Dept., University of Patras, Greece.

Zhang, D. et Lee, W. S. 2004. « Learning to integrate web taxonomies ». *Journal of Web Semantics*. Vol. 2, no 2 (publication numérique disponible à l'adresse www.websemanticsjournal.org/ps/pub/2005-13)

Züll, C., Harkness, J. et Hoffmeyer-Zlotnik, J. H. P. (dir. publ.). 1996. *Text analysis and computers*. Mannheim : ZUMA (ZUMA Nachrichten Spezial).