A

Identifying Authorities in Online Communities

MOHAMED BOUGUESSA, University of Quebec at Montreal
LOTFI BEN ROMDHANE, University of Sousse

Several approaches have been proposed for the problem of identifying authoritative actors in online communities. However, the majority of existing methods suffer from one or more of the following limitations: (1) there is a lack of an automatic mechanism to formally discriminate between authoritative and non-authoritative users. In fact, a common approach to authoritative user identification is to provide a ranked list of users expecting authorities to come first. A major problem of such an approach is where to stop reading the ranked list of users? How many users should be chosen as authoritative? (2) supervised learning approaches for authoritative user identification suffer from their dependency on the training data. The problem here is that labeled samples are more difficult, expensive, and time consuming to obtain than unlabeled ones. (3) several approaches rely on some user parameters to estimate an authority score. Detection accuracy of authoritative users can be seriously affected if incorrect values are used. In this paper, we propose a parameterless mixture model-based approach which is capable of addressing the three aforementioned issues in a single framework. In our approach, we first represent each user with a feature vector composed of information related to its social behaviour and activity in an online community. Next, we propose a statistical framework, based on the multivariate beta mixtures, in order to model the estimated set of feature vectors. The probability density function is therefore estimated and the beta component that corresponds to the most authoritative users is identified. The suitability of the proposed approach is illustrated on real data extracted from Stack Exchange question-answering network and Twitter.

Categories and Subject Descriptors: H.1.2 [**Information Systems**]: User/Machine Systems—*Human factors, Human information processing*; G.3 [**Probability and Statistics**]: Distribution functions, Statistical computing.

General Terms: Algorithms, Design, Experimentation

Additional Key Words and Phrases: Online communities, authoritative users, unsupervised learning, mixture model, multivariate beta

ACM Reference Format:
Mohamed Bouguessa and Lotfi Ben Romdhane, 2014. Identifying Authorities in Online Communities. *ACM Trans. Intell. Syst. Technol.* V, N, Article A (January YYYY), 23 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Online communities have emerged as popular and often effective means that allow people to communicate, share their expertise and experience or perform special roles such as leading or moderating. Nowadays, millions of users exchange information and share knowledge on different kinds of online communities. For example, community question

This work is supported by research grants from the National Sciences and Engineering Research Council of Canada (NSERC).
Author's addresses: M. Bouguessa, Department Computer Science, University of Quebec at Montreal, Montreal, QC, Canada, e-mail: bouguessa.mohamed@uqam.ca;
L. Ben Romdhane, ISITCOM, University of Sousse, email: lotfi.ben.romdhane@usherbrooke.ca.

DOI: http://dx.doi.org/10.1145/0000000.0000000

answering sites such as Stack Exchange Network [1] have become large repositories of valuable knowledge and viable tools for seeking information online. Micro-blogging platforms such as Twitter are another kind of online community that has attracted the attention of many users who share their personal opinions, propagate news or provide the latest promotion information [Kumar et al. 2013], [Yang and Chen-Burger 2012].

Online communities are based on user-centered design which aims to build an interconnected web of users that generate different types of content. The real value of online communities is not only in connecting people to people but also in the quality of content being generated. Users who produce high quality content are known as experts or authoritative [Bonchi et al. 2011], [Pal 2012]. Authoritative users are often seen in the business world as an important source for generating value and motivating contribution. Participants who perform this role are important sources of the value found in online discussion sites. Managers of such web sites would like to be able to identify such authoritative users in order to cultivate valuable online communities.

In fact, authoritative users play a critical role in sustaining and fostering online communities [Bouguessa et al. 2010]. For example, in some online discussion forums and community blogs, reviews of products provided by those influential users can help customers to make better decisions, and businesses to increase sales. In community question answering sites, participants may seek authoritative people as a source of information to complement or replace other sources such as documents and databases in various ways [Bouguessa et al. 2008]. In some cases, managers of online community sites may seek an authoritative user to perform some tasks, e.g. to become a contactor or an employee in order to perform a given organisational role or a social function. This also includes using authoritative people as effective and reliable information filters to select the useful information from the huge mass of information available. To summarize, the characteristics of authoritative users make them the drivers of online communities [Pal et al. 2011]. Identifying and increasing the visibility of such prominent actors improves the experience of other community members and positively impacts the overall value of the service provided.

Identifying authoritative users in a community is mainly related to the problem of expert and influential user identification [Pal et al. 2012]. The problem of identifying such notable actors has been well studied, for which appropriate approaches have been developed. Such approaches can be broadly divided into two categories: ranking-based approaches and attributes-based approaches. Most ranking-based approaches [Kwak et al. 2010], [Weng et al. 2010], [Zhu et al. 2011], [Tang and Yang 2012] calculate some kind of score per user which serves as a measure of the degree of authority. Scores are used in ranking users such that the top K users are considered as authoritative. On the other hand, feature-based approaches [Pal et al. 2011], [Pal and Counts 2011] aim to identify a number of features (generally related to users' social behaviour and activities on the site) which could be potentially used to detect authoritative users. Theses features are then used as attributes of a supervised machine learning process for classifying users as authoritative or not.

In spite of these advanced approaches, identifying authoritative users in online communities continues to pose a challenge to existing algorithms in various ways. In fact, each of the existing approaches suffers from one or even more of the following three shortcomings:

(1) Virtually, ranking-based approaches have so far been only used to obtain a user ranking, expecting authoritative people to come first. On the other hand, less ef-

[1] http://stackexchange.com

fort has been invested on how to automatically discriminate between authoritative and non-authoritative users. In fact, most of the existing ranking-based methods aim to make authoritative users detection more effective in retrieving the top K users only. The weakness of such an approach resides in the unprincipled selection of the value of K. In general, the value of K is often chosen in an *ad hoc* manner. With such an informal approach, it is impossible to be objective or consistent. Furthermore, setting the value of K manually causes practical difficulties in applying ranking-based approaches to real applications, in which prior knowledge about the data under investigation is not always available.

(2) Attributes-based approaches that use supervised machine learning algorithms suffer from their dependency on the training data. In fact, a robust learning approach for authoritative users detection often requires large amounts of labeled training data. The problem here is that labeled samples are more difficult, expensive and time consuming to obtain than unlabeled ones. To alleviate this drawback, semi-supervised approaches could be used with only a small amount of domain knowledge. However, in some applications, domain knowledge in the form of labeled samples is very limited and rarely available. The scarcity of trained data is thus one of the major challenges facing current attributes-based approaches [Bouguessa 2011]. Yet unlabeled data are relatively easy to collect.

(3) A certain number of methods [Tang and Yang 2012], [Agarwal et al. 2012], [Kao et al. 2010], [Zhou et al. 2009] rely on some user-supplied parameters in order to estimate the authority score of the participants. However, in many applications the optimal values of these parameters are difficult to determine. Furthermore, in several cases, a small change in the parameter values influences the final ranking of participants based on the estimated authority scores. This creates a usability problem, as users are not always able to correctly supply the value of the parameters. In some situations, the analyst needs to run the algorithm many times with different parameters to get a feel for which results might be more reasonable. On the other hand, one can argue that for some applications, parameters are a means to incorporate domain knowledge into the process of identifying authoritative actors and thus are beneficial in some situations. However, as mentioned above, domain knowledge is rarely available in real applications. Under this circumstance, we believe that there is a good reason to focus on a parameterless approach that does not require prior knowledge about the data under investigation.

The aforementioned problems have been tackled separately, and specific approaches have been proposed in the literature, which tend to not fit the whole framework well. The main objective of this paper is, instead, to face the three issues in a unified framework. To this end, we propose a simple and systematic, yet effective unsupervised approach for identifying authoritative users in online communities. In a nutshell, in our approach, we first represent each user with a feature vector composed of information related to its social behaviour and activity on the site. Next, we propose a statistical framework, based on the multivariate beta mixtures, in order to model the estimated set of feature vectors. The probability density function is therefore estimated and the beta component that corresponds to the most authoritative users is identified.

We have used the multivariate beta mixtures mainly because the beta distribution offers considerable flexibility and ease of use [Ma 2011], [Ma and Leijon 2009], [Bouguila et al. 2006]. In fact, in contrast with other distributions such as the Gaussian, which permit only a symmetric shape, the beta distribution has a flexible shape; it can be symmetric, asymmetric, or convex. The shapes of the beta distribution are variable enough to allow for an approximation of almost any arbitrary distribution [Bouguila et al. 2006]. It is worth noting that, in contrast to the theoretical work

that exists on the multivariate beta distribution [Olkin and Rubin 1964], very little work has been done on its practical applications [Ma 2011]. The reason that the multivariate beta distribution has not received very much attention might be due to the difficulties involved in parameters estimation, where a closed-form solution does not exist and some approximations are required [Ma 2011]. In this paper, we propose the Expectation-Maximization (EM) algorithm for maximum likelihood estimation of the parameters of the multivariate beta mixture model. To determine the number of components in the mixture we use the integrated classification likelihood Bayesian information criterion.

We summarize the significance of our work as follows:

(1) We view the task of identifying authoritative users from a mixture modeling perspective, based on which we devise an unsupervised approach that is able to automatically discriminate between authoritative and non-authoritative users rather than providing a ranked list of users only. To the best of our knowledge, the work presented in this paper represents the first application of the multivariate beta mixture model to social media data.
(2) The proposed approach is parameterless and does not require any prior knowledge about the data. Furthermore, our method is general in the sense that it can be applied to different kinds of online communities in order to identify authoritative users, while some existing approaches [Weng et al. 2010], [Agarwal et al. 2012], [Liu et al. 2011] deal only with specific types of online applications.
(3) We conducted experiments on data extracted from different online platforms such as Stack Exchange question-answering network and Twitter. The results suggest that the accuracy of our unsupervised approach is comparable (actually, in some cases, even better) to those of attribute-based supervised approaches that have the advantage of using label data.

The remainder of this paper is organized as follows: In Section 2, we provide an overview of related work. Section 3 describes the proposed approach. Section 4 presents experiments and performance results. Finally, our conclusion is given in Section 5.

2. RELATED WORK

In this section, we provide a high level description of recent mainstream approaches for authoritative users identification. With the upsurge of online community sites, the problem of identifying authoritative actors on these sites becomes increasingly important and has been widely investigated recently. However, most of the existing approaches suffer from one or more of the limitations described in Section 1.

In general, ranking-based approaches that use graph algorithms are common in the literature. The key idea here is to represent the environment as a graph in which nodes correspond to users and arcs correspond to the interactions between them. The authority score of nodes is generally measured by means of graph-based ranking algorithms such as PageRank or HITS or their variants. For example, the authors in [Kwak et al. 2010] applied Page-Rank to the Twitter graph in order to identify a ranked list of influential users. Weng et al. [2010] proposed a Page-Rank like algorithm named TwitterRank that uses both the Twitter graph and the processed information from tweets to identify experts in particular topics. In [Romero et al. 2011], the authors designed an algorithm similar to HITS named Influence Passivity algorithm to quantify the influence of users in a Twitter network. This algorithm utilizes both the structural properties of the network as well as the diffusion behaviour among users.

There have also been earlier studies that attempt to identify authoritative actors from Twitter without relying on graph algorithms. For example, Ghosh et al. [2012]

introduced Congos, a system for finding topic experts in Twitter. The proposed system infers first the topical expertise of individual users using Twitter Lists. Then, for a given query topic, it ranks the relative expertise of the users whose topical expertise matches the query. Cha et al. [2010] investigated three measures of influence: indegree (number of people who follow a user), retweets (number of times others forward a user's tweet), and mentions (number of times others mention a user's name) in order to rank users in Twitter.

Authoritative users' identification in online community question answering has also been subject of extensive research. In [Kao et al. 2010], a hybrid method that combines a user's knowledge score and authority score is used to derive a user's expert degree and produce a ranked expert list. It should be noted that this hybrid approach suffers from its dependence on three user-defined parameters, which are used in order to balance the effect of different measures considered in the estimation of the final expertise score of a user. Different values of these parameters may lead to different ranking results. A competition-based method that ranks users in community question-answering sites is proposed in [Liu et al. 2011]. This approach explores pairwise comparisons between users inferred from best answer selections, to estimate a user expertise score. Each pairwise comparison is treated as a two-player competition. From a competition-based perspective, expertise score estimation becomes a related problem to the calculation of the statistical skill rating of players or teams in competitive games.

Some approaches harness topic models techniques in order to identify authoritative users in community question-answering sites. For instance, Zhu et al. [2011] proposed an approach for authority ranking by exploiting the information in both target and relevant categories in community question-answering sites. First, the authors developed a method for measuring the relevancy between categories through topic models. Then, a link analysis approach is proposed for ranking users by considering the information in both target and relevant categories. Riahi et al. [2012] propose to build profiles of users based on their answering history using statistical topic models. Then, these profiles are used in comparison with a newly posted question in order to provide a ranked list of users who are best suited to the posted question. Zhou et al. [2009] developed an approach for finding the top K users for a given question. The proposed approach consists of two components: the expertise model that ranks the users according to their expertise captured by language models and the re-ranking model that re-ranks the candidate users using the question-answer graph. The approach proposed in [Zhou et al. 2009] depends on two user-supplied parameters that are used to smooth the language model, and to specify the reply portion for the hierarchical question-answer thread model.

Identifying authoritative users was not limited to micro-blogging platforms such as Twitter or community question-answering sites. Budalakoti and Bekkerman [2012] studied the LinkedIn social network in order to build a ranked list of users based on estimated authority scores. To this end, the authors in [Budalakoti and Bekkerman 2012] constructed two directed graphs over the same set of users: (1) the invitation graph which is based on invitations to connect, and (2) the navigation graph which is based on users' browsing behaviour. The authority degree of users is then estimated by simultaneously using the authority scores of nodes in one graph to inform the other, and vice versa, in a mutually reinforcing fashion.

Tang and Yang [2012] studied the problem of ranking user influence in online health care community. The proposed approach incorporates users' reply relationships, conversation content and response immediacy to build a weighted social network that represents influence between users. Then, to quantify influence using this weighted network, two ranking approaches are proposed: UserRank and weighted indegree. It is worth noting that the estimation of the weights of the developed social network de-

pends on two user supplied parameters. Experiments in [Tang and Yang 2012] suggest that the variation of the values of these parameters has a substantial impact on the accuracy of results.

Agarwal et al. [2012] studied the problem of identifying a ranked list of influential users in online community blogs. The approach in [Agarwal et al. 2012] is based on the assumption that a blogger can be considered influential if she/he has at least one influential blog post. Starting from this assumption, the authors estimate first the influence of a blog post. Then, the influence of a blogger is calculated using the associated blog post with the maximum influence score. Note that the influence of a blog post is estimated by combining four metrics: (1) inlinks to the blog post, (2) outlinks to other posts, (3) number of comments that a blog post receives, and (4) blog post length. To balance the effect of these four metrics, four user-defined weighting parameters are used in the combination formula. The accuracy of the results depends heavily on proper tuning of the values of these parameters. The experiments in [Agarwal et al. 2012] reveal that changing the values of some of these weights can lead to different ranking results.

Different from the aforementioned methods, attributes-based approaches [Pal 2012] extract a number of characteristics, which could potentially be used to identify authoritative actors. These characteristics are then used as attributes of a machine learning process for classifying users as either authoritative or non-authoritative. For example, Pal et al. [2012] proposed an attribute-based approach for identifying experts and potential experts in community question answering. The authors in [Pal et al. 2012] created several models based on different user features, such as the number of answers, the number of best answers, the number of votes received, etc. Then, a Bagging classifier-learning algorithm was applied to these features models to classify users as authoritative or non-authoritative.

Clustering and ranking algorithms have been also used to identify top authoritative users. In [Pal and Counts 2011] a number of attributes that characterize topical authorities in Twitter are proposed. Based on these attributes a Gaussian mixture-based clustering algorithm is used to group users into two clusters. The main motivation behind dividing users into two Gaussian components is, as suggested in [Pal and Counts 2011], principally to perform ranking on a reduced set of users, instead of the whole set, by selecting the cluster that may contain potential authorities. Once the target cluster is selected, a within-cluster ranking procedure is performed to yield a ranked list of users. The top K users are identified as authoritative. In this setting, it is clear that such an approach inherits the main drawbacks of ranking-based methods: how many users should be chosen as authoritative from a ranked list?

Furthermore, it is worth noting that the use of the Gaussian distribution in [Pal and Counts 2011] appears to be restrictive since this distribution permits symmetric "bell" shape only. However, in many real life applications, the data under investigation are non-Gaussian, that is, skewed data with non-symmetric shapes. In this setting, as observed in [Boutemedjet et al. 2011], the Gaussian distribution may lead to inaccurate modeling (e.g. over estimation of the number of components in the mixture, increase of misclassification errors, etc.). In contrast to the Gaussian model used in [Pal and Counts 2011], in our approach we propose the use of the multivariate beta which permits multiple modes and asymmetry and which can thus approximate a wide variety of shapes. This noticeable feature of the beta distribution enables it to provide an accurate fit of the users feature vectors. Additionally, it is important to note that, in contrast to [Pal and Counts 2011], in our approach we don't impose any restrictions on the number of components in the mixture. As we will show in the next section, the authoritative users component is identified automatically, while the approach proposed in [Pal and Counts 2011], doesn't explicitly identify the component that contains

authoritative users only. All these notable features clearly distinguish the method proposed in this paper from the work described in [Pal and Counts 2011].

Finally, we should point out that, in [Bouguessa et al. 2008], we have developed a probabilistic approach based on the gamma mixture model to identify authoritative users in Yahoo! Answers, a question answering site. It is worth noting that this approach is based only on one feature (the number of best answers), and was evaluated only on data from Yahoo! Answers, so the results are suitable only for this online service. The application of the approach in [Bouguessa et al. 2008] is limited to one-dimensional data and thus could not be used to model the multidimensional user feature vectors considered in this study.

In the following section, we propose a more general approach for identifying authorities in online communities. In contrast to our previous work [Bouguessa et al. 2008], the approach proposed in this paper explores several features that may characterize authoritative users and uses a more flexible model than the gamma mixture to discriminate between authoritative and non-authoritative users. In fact, in contrast to the gamma distribution, the beta distribution is very versatile and it is capable of modelling a variety of uncertainties. Specifically, the gamma distribution may be L-shaped, skewed to the right, or symmetric, while the beta distribution may be L-shaped, U-shaped, J-shaped, skewed to the left, skewed to the right, or symmetric [Bouguila et al. 2006]. The shapes of Gaussian, gamma and uniform distributions are special cases of the beta distribution [Boutemedjet et al. 2011]. This great shape flexibility of the beta distribution provides a better fitting of the data under investigation which leads, in turn, to a substantially improved modeling accuracy.

3. PROPOSED APPROACH

Let $\mathbf{U} = \{U_1, \ldots, U_N\}$ denotes the set of N users such that each user U_i is represented by D-dimensional feature vector $\vec{X_i} = (x_{i1}, \ldots, x_{iD})^T$. Each element $x_{id},\ (i = 1, \ldots, N;\ d = 1, \ldots, D)$ of the vector $\vec{X_i}$ corresponds to a value that would reflect the authority of a user in a specific online community. We assume that higher feature values indicate a high level of authority, while the smallest ones correspond to non-authoritative users. Note that, in our method, we consider different features that may characterize authoritative actors. For example, to identify authorities in community question answering, we will consider features such as the number of answers, the number of best answers, the number of votes received, etc. It is clear that the values of these features may have different scales, and it makes sense to perform some transformation on these values.

In our approach, we first perform log-transformation to all the estimated feature values of all users. Such log-transformation aims to squeeze together the large values that characterize authoritative actors and stretch out the smallest values, which correspond to non-authoritative users. This squeezing and stretching yields comparable feature values and also contributes to enhancing the contrast between largest and smallest values. Then, we normalize the log-transformed values of each feature in the interval [0,1]. As a result, without loss of generality, all transformed feature values will have comparable normalized values. In the remainder of this paper, we use only the normalized values of the users' feature vectors $\{\vec{X_i}\}$.

Finally, based on the normalized feature vectors, we propose a statistical approach that uses the multivariate beta mixture model to automatically discriminate authoritative from non-authoritative users. Specifically, $\{\vec{X_i}\}$ can be considered as coming from several underlying probability distributions. Each distribution is a component of the multivariate beta mixture model that represents a set of users' feature vectors which are close one to another, and all the components are combined by a mixture

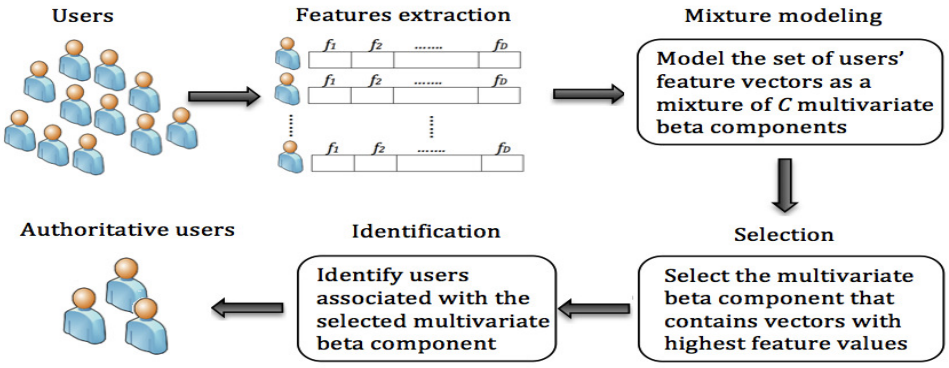


Fig. 1: Workflow of the proposed approach.

form. The component which contains vectors with the highest score values corresponds to authoritative users. Figure 1 provides a simple visual illustration of the proposed approach.

3.1. The Multivariate Beta Mixture Model

In this paper, we expect that the normalized user feature vector $\vec{X_i} = (x_{i1}, \ldots, x_{iD})^T$ follow a mixture density of the form

$$\mathcal{F}(\vec{X_i}|\alpha, \vec{a}, \vec{b}) = \sum_{c=1}^{C} \alpha_c \, \mathcal{F}_c(\vec{X_i}|\vec{a_c}, \vec{b_c}) \tag{1}$$

where $\mathcal{F}_c$ is the cth multivariate beta distribution; C denotes the number of components in the mixture; $\vec{a} = \{\vec{a_1}, \ldots, \vec{a_C}\}$ and $\vec{b} = \{\vec{b_1}, \ldots, \vec{b_C}\}$. $\vec{a_c}$ and $\vec{b_c}$ are the parameters of the cth component with $\vec{a_c} = (a_{c1}, \ldots, a_{cD})^T$ and $\vec{b_c} = (b_{c1}, \ldots, b_{cD})^T$. $\alpha = \{\alpha_1, \ldots, \alpha_C\}$ represents the mixing coefficients such that $\sum_{c=1}^{C} \alpha_c = 1$ and $\alpha_c > 0$.

The multivariate beta distribution can be obtained by cascading a set of beta variables together, that is, each element in the D-dimensional vector $\vec{X_i}$ is a scalar beta variable [Ma 2011], [Ma and Leijon 2009]. The probability density function of the cth multivariate beta component is expressed as

$$\mathcal{F}_c(\vec{X_i}|\vec{a_c}, \vec{b_c}) = \prod_{d=1}^{D} f(x_{id}|a_{cd}, b_{cd}) \tag{2}$$

$f(x_{id}|a_{cd}, b_{cd})$ is the probability density function of the univariate beta distribution which is given by

$$f(x_{id}|a_{cd}, b_{cd}) = \frac{\Gamma(a_{cd} + b_{cd})}{\Gamma(a_{cd})\Gamma(b_{cd})} x_{id}^{a_{cd}-1}(1 - x_{id})^{b_{cd}-1} \tag{3}$$

where $\Gamma(.)$ is the gamma function given by $\Gamma(y) = \int_0^\infty t^{y-1} \exp(-t)dt; t > 0$.

3.2. Maximum Likelihood for the Multivariate Beta Mixture

The maximum likelihood estimation approach can be used to find the parameters of the mixture model. Let $\Theta = \{\alpha_1, \ldots, \alpha_C, \vec{a_1}, \ldots, \vec{a_C}, \vec{b_1}, \ldots, \vec{b_C}\}$ denotes the set of un-

known parameters of the mixture and $\mathbf{X} = \{\vec{X_1}, \ldots, \vec{X_N}\}$ the set of the normalized users feature vectors. The likelihood function corresponding to C components is defined as

$$Ł(\mathbf{X}|\Theta) = \prod_{i=1}^{N} \mathcal{F}(\vec{X_i}|\alpha, \vec{a}, \vec{b}) = \prod_{i=1}^{N} \sum_{c=1}^{C} \alpha_c \, \mathcal{F}_c(\vec{X_i}|\vec{a_c}, \vec{b_c}) \tag{4}$$

The maximum likelihood of the mixture parameters can be estimated using the Expectation Maximization (EM) algorithm [Dempster et al. 1977]. Accordingly, for each $\vec{X_i}$, we assign a *C*-dimensional indication vector $\vec{Z_i} = (z_{i1}, \ldots, z_{iC})^T$ such that

$$z_{ic} = \begin{cases} 1 \;\; if \;\; \vec{X_i} \; belongs \; to \; component \; c \\ 0 \qquad\qquad otherwise \end{cases} \tag{5}$$

Let $\mathbf{Z} = \{\vec{Z_1}, \ldots, \vec{Z_N}\}$ denotes the set of indication vectors. The likelihood function of the complete data is given by

$$Ł(\mathbf{X}, \mathbf{Z}|\Theta) \; = \; \prod_{i=1}^{N} \prod_{c=1}^{C} \left[\alpha_c \, \mathcal{F}_c(\vec{X_i}|\vec{a_c}, \vec{b_c})\right]^{z_{ic}} \tag{6}$$

Consequently, the logarithm of the likelihood is given by

$$\begin{aligned} log(Ł(\mathbf{X}, \mathbf{Z}|\Theta)) \; &= \; \sum_{i=1}^{N} \sum_{c=1}^{C} z_{ic} \, log\big[\alpha_c \, \mathcal{F}_c(\vec{X_i}|\vec{a_c}, \vec{b_c})\big] \\ &= \; \sum_{i=1}^{N} \sum_{c=1}^{C} z_{ic} \, log\big[\alpha_c \prod_{d=1}^{D} f(x_{id}|a_{cd}, b_{cd})\big] \\ &= \; \sum_{i=1}^{N} \sum_{c=1}^{C} z_{ic} \big[log(\alpha_c) + \sum_{d=1}^{D} log(f(x_{id}|a_{cd}, b_{cd}))\big] \end{aligned} \tag{7}$$

The EM algorithm can be used to estimate Θ. Specifically, the algorithm iterates between an Expectation step and an Maximization step in order to produce a sequence estimate $\{\hat{\Theta}\}^{(I)}, (I = 0, 1, 2, \ldots)$, where I denotes the current iteration step, until the change in the value of the log-likelihood in (7) is negligible. Details of each step are given bellow.

In the Expectation step: each latent variable z_{ic} is replaced by its expectation as follows

$$\hat{z}_{ic}^{(I)} = E[z_{ic}|X_i, \Theta] = \frac{\hat{\alpha}_c^{(I)} \mathcal{F}_c(\vec{X_i}|\widehat{\vec{a_c}}, \widehat{\vec{b_c}})}{\sum_{k=1}^{C} \hat{\alpha}_k^{(I)} \mathcal{F}_k(\vec{X_i}|\widehat{\vec{a_k}}, \widehat{\vec{b_k}})} \tag{8}$$

In the Maximization step: the mixing coefficients $\{\alpha_c\}$ and the parameters $\{\vec{a_1}, \ldots, \vec{a_C}, \vec{b_1}, \ldots, \vec{b_C}\}$ are calculated using the values of $\widehat{z}_{ic}$ estimated in the Expectation step. Specifically, the mixing coefficients are calculated as

$$\hat{\alpha}_c^{(I+1)} = \frac{\sum_{i=1}^{N} \hat{z}_{ic}^{(I)}}{N},\ c = 1, \ldots, C \tag{9}$$

Let us now focus on estimating the parameters $\{\vec{a_c} = (a_{c1}, \ldots, a_{cD})^T\}_{(c=1,\ldots,C)}$ and $\{\vec{b_c} = (b_{c1}, \ldots, b_{cD})^T\}_{(c=1,\ldots,C)}$. We note that the parameters pair $\{a_{cd}, b_{cd}\}$ is independent from all other pairs [Ma and Leijon 2009]. The problem of estimating the parameters of the model can thus be reduced to the estimation of the parameters pair $\{a_{cd}, b_{cd}\}$ over each dimension of the data under investigation. In this setting, the value $\{\widehat{a}_{cd}, \widehat{b}_{cd}\}$ that maximize the likelihood can be obtained by taking the derivative of the expectation of the log-likelihood of the complete data with respect to a_{cd} and b_{cd} and setting the gradient equal to zero as

$$\begin{bmatrix} \frac{\partial E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{\partial a_{cd}} \\ \\ \frac{\partial E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{\partial b_{cd}} \end{bmatrix} = 0 \tag{10}$$

where

$$\begin{aligned} \frac{\partial E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{\partial a_{cd}} &= \sum_{i=1}^{N} \widehat{z}_{ic}\ \frac{\partial}{\partial a_{cd}} log(f(x_{id}|a_{cd}, b_{cd})) \\ &= \sum_{i=1}^{N} \widehat{z}_{ic} \Big[\frac{\partial}{\partial a_{cd}} log\Big(\frac{\Gamma(a_{cd}+b_{cd})}{\Gamma(a_{cd})\Gamma(b_{cd})} x_{id}^{a_{cd}-1}(1-x_{id})^{b_{cd}-1}\Big)\Big] \\ &= \sum_{i=1}^{N} \widehat{z}_{ic} \Big[\frac{\Gamma'(a_{cd}+b_{cd})}{\Gamma(a_{cd}+b_{cd})} - \frac{\Gamma'(a_{cd})}{\Gamma(a_{cd})} + log(x_{id})\Big]. \end{aligned} \tag{11}$$

and

$$\begin{aligned} \frac{\partial E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{\partial b_{cd}} &= \sum_{i=1}^{N} \widehat{z}_{ic}\ \frac{\partial}{\partial b_{cd}} log(f(x_{id}|a_{cd}, b_{cd})) \\ &= \sum_{i=1}^{N} \widehat{z}_{ic} \Big[\frac{\partial}{\partial b_{cd}} log\Big(\frac{\Gamma(a_{cd}+b_{cd})}{\Gamma(a_{cd})\Gamma(b_{cd})} x_{id}^{a_{cd}-1}(1-x_{id})^{b_{cd}-1}\Big)\Big] \\ &= \sum_{i=1}^{N} \widehat{z}_{ic} \Big[\frac{\Gamma'(a_{cd}+b_{cd})}{\Gamma(a_{cd}+b_{cd})} - \frac{\Gamma'(b_{cd})}{\Gamma(b_{cd})} + log(1-x_{id})\Big]. \end{aligned} \tag{12}$$

Equations (10), (11) and in (12) yield the following expression

$$\begin{bmatrix} \sum_{i=1}^{N} \widehat{z}_{ic}\big[\psi(a_{cd}+b_{cd}) - \psi(a_{cd}) + log(x_{id})\big] \\ \\ \sum_{i=1}^{N} \widehat{z}_{ic}\big[\psi(a_{cd}+b_{cd}) - \psi(b_{cd}) + log(1-x_{id})\big] \end{bmatrix} = 0 \tag{13}$$

where $\psi(.)$ is the digamma function given by $\psi(\lambda) = \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}$.

Since the digamma function is defined though an integration, a closed-form solution to (13) does not exists [Ma 2011]. So the parameters pair $\{a_{cd}, b_{cd}\}$ can be estimated using the Newton-Raphson, a tangent method for root finding. Specifically, $\{a_{cd}, b_{cd}\}$ are estimated iteratively:

$$\begin{bmatrix} a_{cd}^{(I+1)} \\ \\ b_{cd}^{(I+1)} \end{bmatrix} = \begin{bmatrix} a_{cd}^{(I)} \\ \\ b_{cd}^{(I)} \end{bmatrix} - \begin{bmatrix} \frac{\partial E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{\partial a_{cd}} \\ \\ \frac{\partial E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{\partial b_{cd}} \end{bmatrix} \times \begin{bmatrix} \frac{\partial^2 E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{(\partial a_{cd})^2} & \frac{\partial^2 E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{\partial a_{cd}\,\partial b_{cd}} \\ \\ \frac{\partial^2 E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{\partial b_{cd}\,\partial a_{cd}} & \frac{\partial^2 E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{(\partial b_{cd})^2} \end{bmatrix}^{-1} \tag{14}$$

where

$$\begin{aligned} \frac{\partial^2 E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{(\partial a_{cd})^2} &= \sum_{i=1}^{N} \widehat{z}_{ic}\big[\psi'(a_{cd}+b_{cd}) - \psi'(a_{cd})\big], \\ \frac{\partial^2 E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{\partial a_{cd}\,\partial b_{cd}} &= \sum_{i=1}^{N} \widehat{z}_{ic}\big[\psi'(a_{cd}+b_{cd})\big], \\ \frac{\partial^2 E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{\partial b_{cd}\,\partial a_{cd}} &= \sum_{i=1}^{N} \widehat{z}_{ic}\big[\psi'(a_{cd}+b_{cd})\big], \\ \frac{\partial^2 E[log(Ł(\mathbf{X},\mathbf{Z}|\Theta))]}{(\partial b_{cd})^2} &= \sum_{i=1}^{N} \widehat{z}_{ic}\big[\psi'(a_{cd}+b_{cd}) - \psi'(b_{cd})\big]. \end{aligned}$$

where $\psi'(.)$ is the trigamma function.

The iterative formula of the Newton-Raphson algorithm expressed by (14) requires starting values for $\{a_{cd}^{(0)}, b_{cd}^{(0)}\}$. In our implementation, we have used the method of moments estimators of the beta distribution to define these initial values. Specifically, the moments estimators of $a_{cd}^{(0)}$ and $b_{cd}^{(0)}$ are

$$\begin{aligned} \hat{a}_{cd}^{(0)} &= \overline{\mu}_{cd}\left[\frac{\overline{\mu}_{cd}(1-\overline{\mu}_{cd})}{\sigma_{cd}^2} - 1\right], \\ \hat{b}_{cd}^{(0)} &= (1-\overline{\mu}_{cd})\left[\frac{\overline{\mu}_{cd}(1-\overline{\mu}_{cd})}{\sigma_{cd}^2} - 1\right]. \end{aligned} \tag{15}$$

where $\overline{\mu}_{cd}$ and σ_{cd}^2 denotes respectively the sample mean and variance of the the data values belonging to the cth component which are projected along dimension d. Note that the Newton-Raphson algorithm converges, as our estimate of a_{cd} and b_{cd} change by less than a small positive value ξ with each successive iteration, to $\widehat{a}_{cd}$ and $\widehat{b}_{cd}$.

The EM algorithm can now be used to estimate the maximum likelihood of the distribution parameters. Note that EM is highly dependent on initialization [Figueiredo and Jain 2002]. To alleviate this problem, a common solution is to perform initialization by mean of clustering algorithms. For this purpose we first implement the Fuzzy C-Means (FCM) algorithm [Bezdek 1981] in order to partition the set $\{\vec{X}_i\}_{i=1...,N}$ into C components. Then, based on such partition, we estimate the parameters of each

ALGORITHM 1: Estimating the number of components in the mixture

Input : $\{\vec{X}_i\}_{(i=1,\ldots,N)}$, C_max
Output: The optimal number of components C
begin
  for $C = 1$ **to** C_max **do**
    if *C==1* **then**
      Estimate $\{\widehat{a}_d, \widehat{b}_d\}_{d=1,\ldots,D}$ based on (14);
      Compute the value of ICL-BIC(C) using (17);
    else
      Apply the FCM algorithm as an initialization of the EM algorithm;
      Estimate the mixture parameters by alternating the following two steps:
      — **E-Step**: Compute $\hat{z}_{ic}^{(I)}$ using (8);
      — **M-Step**:
        (1) Estimate the mixing coefficients using (9);
        (2) Estimate $\{\widehat{a}_{cd}, \widehat{b}_{cd}\}_{(c=1,\ldots,C;\ d=1,\ldots,D)}$ using (14);
      Repeat E-Step and M-Step until the change in (7) is negligible;
      Compute the value of ICL-BIC(C) using (17);
    end
  end
  Select $\hat{C}$, such that $\hat{C} = arg\ min_C\Big\{\text{ICL-BIC}(C),\ C = 1, \ldots, C\_max\Big\}$;
end

component using the method of moment estimator of the beta distribution [Bain and Engelhardt 2000] and set them are as initial parameters to the EM algorithm.

3.3. Estimating the Number of Components in the Mixture

The number of components C in the mixture is an unknown parameter that must be estimated. Several model selection approaches have been proposed to estimate C. In this paper, we implemented a deterministic approach that use the EM algorithm in order to obtain a set of candidate models for the range value of C (from 1 to C_max, the maximal number of components in the mixture) which is assumed to contain the optimal C [Figueiredo and Jain 2002]. The number of components is the selected according to

$$\hat{C} = arg\ min_C\Big\{MC\big(\hat{\Theta}(C), C\big), C = 1, \ldots, C\_max\Big\} \tag{16}$$

where $MC\big(\hat{\Theta}(C), C\big)$ is some model selection criterion, and $\hat{\Theta}(C)$ is an estimate of the mixture parameters assuming that it has C components. In this paper, we use the integrated classification likelihood Bayesian information criterion (ICL-BIC) [Ji et al. 2005] which is given by

$$\text{ICL} - \text{BIC}(C) = -2\log(\text{Ł}_C) + p\log(N) - 2\sum_{i=1}^{N}\sum_{c=1}^{C}\hat{z}_{ic}log(\hat{z}_{ic}) \tag{17}$$

where Ł_C is the logarithm of the likelihood at the maximum likelihood solution for the investigated mixture model, and p is the number of parameters estimated. The procedure for estimating the number of components in the mixture is summarized in Algorithm 1.

ALGORITHM 2: Authoritative users identification procedure

Input : A set $U = \{U_1, \ldots, U_N\}$ of N users
Output: A set $A = \{A_1, \ldots, A_K\}$ of K authoritative users
begin
For a given online community, estimate a feature vector $\vec{X}_i$ for each user;
Normalize $\{\vec{X}_i\}$, as discussed at the beginning of Section 3;
Apply Algorithm 1 to cluster the users into C multivariate beta components;
Use the results of the EM algorithm to decide about the membership of $\vec{X}_i$ in each component;
Select the multivariate beta component that corresponds to the highest feature values;
Identify authoritative users in U associated with the set of $\vec{X}_i$ that belong to the selected component and store them in A;
Return A;
end

3.4. Automatic Identification of Authoritative Users

Once the optimal number of components have been identified, we focus now on detecting the multivariate beta component that corresponds to authoritative users. To this end, we used the results of the EM algorithm in order to derive a classification decision about which users feature vector $\vec{X}_i$ belongs to which component in the mixture. In fact, the EM algorithm yields the final estimated posterior probability $\hat{z}_{ic}$, the value of which represents the posterior probability that $\vec{X}_i$ belongs to component c. We assign $\vec{X}_i$ to the component that corresponds to the maximum value of $\hat{z}_{ic}$. We thus divide the set of user feature vectors into several components. As discussed earlier, in our approach we assume that authoritative users are characterized by high feature values [2]. Therefore, we are interested by the multivariate beta component which contains vectors with the highest values. To identify such a component, we first compute, for each component in the mixture, the average of the projected feature values along each dimension. Then, we select the component with the larger average value as our target component. Accordingly, users associated with the set of $\vec{X}_i$ that belong to such a component correspond to authoritative users. The steps described in Algorithm 2 summarize our authoritative users identification procedure.

4. EMPIRICAL EVALUATION

Our goal now is to illustrate the effectiveness of the proposed approach to identify authoritative users in different online community sites. To this end, we put our approach to work using data sets collected from: (1) Stack Exchange, a fast growing network of many question and answer sites, and (2) Twitter, a popular microblogging platform. The data collected from each specific online service are used to build one feature vector per user. Each vector is composed of a set of information that reflects the user authority in a specific online community. These feature vectors are then used as input to the proposed approach to identify authoritative actors. It is important to note that, in this paper, we do not aim at defining new features to identify authoritative users. This would be far beyond the scope of this study. In our experiments, we utilize existing features that may characterize authoritative actors and we focus on illustrating the suitability of the proposed approach to discriminate between authoritative and non-authoritative users.

[2]Note that in the case where an authoritative user is characterized by small feature values, we simply perform linear inversion so that high values will correspond to authorities.

Note that, for the purpose of evaluation, we constructed labelled collections of users by manually classifying each user as either authoritative or non-authoritative. These labelled samples are used as ground truth to evaluate the efficacy of the proposed approach. To this end, we used the following standard metrics: (1) Accuracy, which corresponds to the proportion of correctly partitioned users, (2) Correct Detection (CD) rate, measuring the proportion of authoritative users that are correctly identified as authoritative, (3) False Alarm (FA) rate, corresponding to the proportion of non authoritative users incorrectly classified as authoritative, and (4) F-measure of the authoritative users' class, corresponding to the harmonic mean between precision and recall of the authoritative user class.

To demonstrate the capability of our approach, we compared it with various attributes-based learning algorithms belonging to three different categories of classifiers: (1) Meta classifiers, (2) Tree classifiers and (3) Function-based classifiers. For meta classifiers, we considered AdaBoost, Bagging, Decorate, LogitBoost and MultiBoost. For the Tree classifier camp, we choose ADTree and Random Forest. Finally, we selected RBF Networks and SVM as representative of function-based classifiers. We believe our choice of algorithms covers a wide spectrum of attributes-based learning approaches. In our experiments, we used Weka [3] for running these algorithms. The classification experiments were performed using 10-fold cross-validation to improve the reliability of classifier evaluation.

In the following, we illustrate the suitability of our approach to identify authoritative users in online community question answering, using data extracted from Stack Exchange. Further, we also demonstrate the effectiveness of the proposed approach to identify topical authorities in microblogs using data extracted from Twitter.

4.1. Identifying Authorities in Community Question Answering

Community question answering refers to online services where users come to ask and answer questions and share their knowledge. In this paper, we are interested by focused question answering sites in which users ask concrete questions and expect factual answers, ideally given by an authoritative user who has deep expertise in the domain area. Stack Exchange, seem well suited for the task of our study, as it contain a number of question answering web sites in which knowledge sharing and factual expertise are sought. In our experiments, we analysed data from two different web sites: (1) Game Development [4], a question answering site for professional and independent game developers, and (2) Unix & Linux [5], a question answering forum for users of Linux, FreeBSD and other Unix-like operating systems. Both web sites discourage subjective or argumentative questions [6] and are built around technically focused communities which favour knowledge sharing and factual expertise [Pal et al. 2012].

We used Stack Exchange Data Explorer [7], a web tool for querying data from the Stack Exchange network, to collect data from Game Development and Unix & Linux forums. Specifically, for each web site, we collected data that represented users' activities between January 2013 and Jun 2013. For the purpose of evaluation, three human annotators were recruited to analyze the collected data in order to produce labeled collections of authoritative and non-authoritative users. Similar to the work in [Pal et al. 2012], the labeling was done by looking at each user's profile page available in Game Development and in Unix & Linux forum, and also by examining the latest 10

[3] http://www.cs.waikato.ac.nz/ml/weka/
[4] http://gamedev.stackexchange.com
[5] http://unix.stackexchange.com
[6] http://stackexchange.com/about
[7] http://data.stackexchange.com

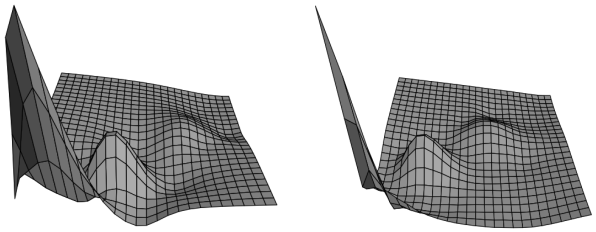
answers provided by each user. As a result of this process, two hand-coded data sets were created. The first data set contains 1000 Game Development users; 103 users (10.3% of the whole data set) among them were labeled as authoritative and 897 users were labelled as non-authoritative. The second data set contains 1000 Unix & Linux users, out of which 98 users (9.8% of the whole data set) were labeled as authoritative and the rest (902 users) as non-authoritative.

Next, we represented each user as a feature vector such that each element of the vector contains information that would reflect the authority of a user in an online community. In this paper, for each user, we considered the following four representative user features: (1) number of answers, (2) number of best answers, (3) number of votes received and (4) Z-score = $(a-q)/\sqrt{(a+q)}$, where a is the number of answers provided by the user and q is the number of questions asked by that user. Higher values of these features indicates a high level of authority while the smallest ones correspond to non-authoritative users. In fact, in the context of specialized technical question answering services and in comparison to non-authoritative users, authoritative users tend to answer a high number of questions; a significant fraction among them are rewarded as best answers. The vote given by participants is also a potential indicator of authoritativeness since it reflects the satisfaction of community members with the provided answers. Finally, the Z-score is a reliable measure that helps to distinguish authoritative from non-authoritative users. It has been reported in [Zhang et al. 2007] that the Z-score measure performs better than the graph-based approach such as Page Rank, HITS and some of their variants, for finding authorities in question answering services. Authoritative users tend to have high Z-score values since, in general, they answer more than they ask [Pal et al. 2012], [Zhang et al. 2007].

Experiment 1. The goal of this first set of experiments is to evaluate the detection accuracy of our approach using different subsets of the four user features (number of answers, number of best answers, number of votes received and Z-score) considered in this study. To this end, using the data collected from Game Development and Unix & Linux web sites, we created, for each online service, several data sets using the following subsets of features: (1) number of answers and number of best answers, (2) number of best answers and Z-score, (3) number of best answers and number of votes received, (4) number of answers and Z-score, (5) number of answers, number of best answers and Z-score, (6) number of answers, number of best answers and number of votes received, (7) number of best answers, number of votes received and Z-score, and, finally, (8) the four user features. Then, for each constructed data set, we used our approach to identify authoritative users. To this end, we set C_max to 6 [8] in all our experiments and selected the optimal number of components that minimize ICL-BIC. We found that the number of components varies from three to four. For the purpose of illustration and in order to not encumber the paper, we show in Figure 2 the estimated probability density function of the users' vectors resulting from several $2D$ features combinations over Unix & Linux data only. The multivariate beta component that represents the highest feature values corresponds to authoritative users.

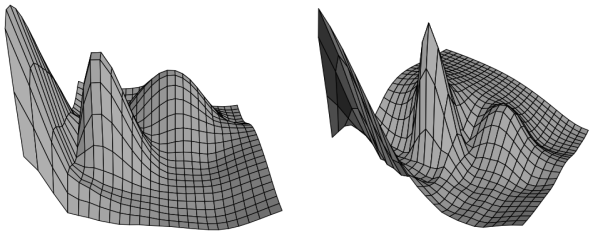
The created data sets differed only in the underlying features used but have the same class labels that designate authoritative and non-authoritative users. Obviously, we have ignored these class labels when applying our approach but we used them to evaluate the detection accuracy of the proposed method. Figure 3 illustrates the results, evaluated with Accuracy, CD rate, FA rate and F-measure, of different combinations of features for both Game Development and Unix & Linux data. Shaded regions

[8]The reader should be aware that the choice of C_max is not limited to 6 and the user can choose other values.



(a) Number of answers and number of best answers.

(b) Number of best answers and Z-score.



(c) Number of best answers and number of votes received.

(d) Number of answers and Z-score.

Fig. 2: Density curves of several $2D$ user features combinations over Unix & Linux data.

Input features	Accuracy	CD	FA	F-measure
# of answers and # of best answers	97.6%	76.6%	0.0%	0.868
# of best answers and Z-score	95.7%	58.2%	0.0%	0.736
# of best answers and # of votes received	90.7%	89.3%	9.1%	0.664
# of answers and Z-score	95.3%	100%	5.2%	0.814
# of answers, # of best answers and Z-score	97.7%	78.6%	0.1%	0.875
# of answers, # of best answers and # of votes received	92.9%	100%	7.9%	0.743
# of best answers, # of votes received and Z-score	92.1%	100%	9.8%	0.700
All features	**99.1%**	**97.0%**	**0.6%**	**0.956**

(a)

Input features	Accuracy	CD	FA	F-measure
# of answers and # of best answers	97.6%	100%	2.6%	0.890
# of best answers and Z-score	92.7%	94.8%	7.5%	0.718
# of best answers and # of votes received	93.6%	92.8%	6.3%	0.739
# of answers and Z-score	89.1%	96.9%	11.7%	0.635
# of answers, # of best answers and Z-score	98.4%	94.8%	1.2%	0.920
# of answers, # of best answers and # of votes received	98.9%	93.9%	0.5%	0.943
# of best answers, # of votes received and Z-score	96.6%	92.8%	2.9%	0.842
All features	**99.2%**	**95.9%**	**0.4%**	**0.959**

(b)

Fig. 3: Performance results of the proposed approach over : (a) Game Development data, (b) Unix & Linux data.

in this figure correspond to the best values of the four evaluation metrics considered in the experiment.

As depicted by Figure 3, the use of the four user features, namely number of answers, number of best answers, number of votes received, and Z-score, yields the highest accuracy and F-measure values for Game Development and Unix & Linux data. In

fact, using the four input features, our approach achieves an accuracy grater than 99% and F-measure over 0.95, both pointing to accurate results. The use of the four user features also yields high CD rates (97% and 95.9%) and low FA rates (0.6% and 0.4%) suggesting their practical usability to accurately identify authoritative users.

On the other hand, we observe that, for some features combinations, our approach is able to correctly identify all authoritative users with the expense of also selecting a number of non-authoritative users as authoritative. For example, as can be seen from Figure 3(a), for three different features combinations, our approach reports CD rates of 100% with a relatively high FA rate values (5.2%, 7.9% and 9.8%). For Unix & Linux (Figure 3(b)), when the number of answers and the number of best answers are used, our approach achieves a CD rate of 100% and FA rate of 2.6%. The second best CD rate (96.9%) is achieved when the number of answers and Z-score are considered as input features to the proposed approach. However, using the same subset of features, the FA rate is 11.7% which is the highest value compared to other values reported by the proposed approach using different features combinations. For Game Development (Figure 3(a)), the lowest FA rate (0%) is achieved when: (1) the number of answers and number of best answers, and (2) number of best answers and Z-scores are used as input features. However, using the same subsets of features, the CD rate is relatively low (76.6% and 58.2%). On the other hand, for Unix & Linux (Figure 3(b)), the lowest FA rate (0.4%) is achieved when using all the four user features.

Overall, this experiment seem to suggest that, in general, a substantial improvement is gained in identifying authoritative users by considering the number of answers, number of best answers, number of votes received, and Z-score together to discover authorities. In fact, as can be seen from Figure 3, the combination of these four user features yields the best trade-off between CD rate and FA rate to get higher F-measure for both Game Development and Unix & Linux data.

Experiment 2. The goal of this second set of experiments is to compare the performance of our approach to several machine learning algorithms. Note that, in this experiment, we have used all the four user features that characterize authoritative users as input to all competing algorithms. Figure 4 illustrates the results of the compared algorithms. Shaded regions in this figure correspond to the best values of Accuracy, CD rate, FA rate and F-measure.

Figure 4 shows that, in general, most competing algorithms were fairly accurate for both data sets. For Game Development (Figure 4(a)), the proposed approach and ADTree report the highest Accuracy, CD rate, FA rate and F-measure. In fact, our approach and ADTree achieve an Accuracy of 99.1%, CD and FA rates of 97% and 0.6% respectively and finally an F-measure of 0.956, all pointing to accurate results. As depicted by Figure 4(a), other approaches, such as AdaBoost, Bagging and LogitBoost, provide quite similar results to that of our method and ADTree. Results in Figure 4(a) also suggest that Random Forest and SVM show good performance. Finally, Decorate and RBF Network, although they were not as successful as the proposed approach and the remaining learning algorithms in detecting authoritative users, provide also acceptable results.

For Unix & Linux (Figure 4(b)), the best F-measure value is achieved by AdaBoost (0.951) followed by our approach (0.943). The proposed method reports the lowest FA rate (0.5%) among all competing approaches and the highest accuracy (98.9%) together with AdaBoost. In terms of CD rate, our approach achieves 93.9% while six of the eight learning algorithms considered in the comparison achieve a higher CD rate (96.1%). On the other hand, results in Figure 4(b) suggest that Decorate and Random Forest provide good results while the performance of Bagging and RBF Network is less competitive.

Algorithm	Accuracy	CD	FA	F-measure
Proposed	**99.1%**	**97%**	**0.6%**	**0.956**
AdaBoost	99.1%	96.1%	0.6%	0.956
Bagging	99%	96.1%	0.7%	0.952
Decorate	98.4%	92.2%	0.9%	0.922
LogitBoost	99%	96.1%	0.7%	0.952
MultiBoostAB	98.9%	97%	0.9%	0.948
ADTree	99.1%	97%	0.6%	0.956
Random Forest	98.9%	96.1%	0.8%	0.947
RBF Network	98.3%	94.2%	1.2%	0.919
SVM	98.9%	95.1%	0.8%	0.942

(a)

Algorithm	Accuracy	CD	FA	F-measure
Proposed	**98.9%**	**93.9%**	**0.5%**	**0.943**
AdaBoost	98.9%	96.1%	0.7%	0.951
Bagging	97.9%	92.2%	1.3%	0.904
Decorate	98.3%	96.1%	1.3%	0.925
LogitBoost	98.7%	96.1%	0.9%	0.942
MultiBoostAB	97.7%	94.1%	1.8%	0.897
ADTree	98.7%	94.1%	0.7%	0.941
Random Forest	98.3%	96.1%	1.3%	0.925
RBF Network	98.1%	96.1%	1.6%	0.916
SVM	98.7%	96.1%	0.9%	0.942

(b)

Fig. 4: Accuracies of compared algorithms on: (a) Game Development data, (b) Unix & Linux data.

4.2. Identifying Topical Authorities in Microblogs

Microblogging is an emerging and important platform for exchanging real-time information on the Web [Cheng et al. 2013]. One of the most notable microblogging sites is Twitter [Kumar et al. 2013]. Users of Twitter post message updates, called tweets, of up to 140 characters. A hash symbol (#), called a hash-tag, is associated to each tweet. This symbol is usually employed to mark keywords or topics [Pal and Counts 2011]. Twitter users follow others, or are followed. The relationship of following and being followed requires no reciprocation. A Twitter user is allowed to choose who she/he wants to follow without seeking any permission. Conversely, a user may also be followed by other people without granting any permission. Being a follower on Twitter means that the user receives all the tweets from those the user follows. A user can also propagate interesting tweets, originally posted by someone else, to the user's followers. This act is commonly called retweeting and the tweets resulting from this act are preceded by "RT @username". Finally, a user can also mention other users using the "@username" tag. As suggested in [Cha et al. 2010], this act is called mentioning.

Twitter publishes a lot of tweets per second. These tweets contain a wide variety of information, ranging from conversational tweets to highly relevant information on specific topics [Bigonha et al. 2012]. As a result, Twitter has become one of the most important mediums of communication. This success can be attributed to the significant number of participants, with different skills and expertise, who post messages related to a wide variety of topics. This makes Twitter an interesting case study, since it contains a rich store of information. In this context, automatically identifying the users that are recognized sources of relevant and trustworthy information on specific topics is crucial. In fact, such topical authoritative users drive the communication and may influence other tweeters.

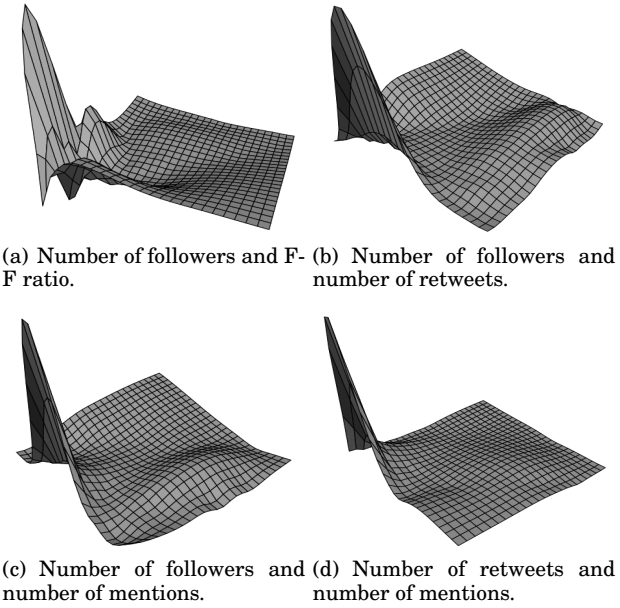
We used the search API of Twitter to extract data in order to evaluate our approach to identify authoritative actors in a topic-based scenario. The collected data set regards the 2012 Quebec provincial election. The data set consists of tweets posted between Au-

gust 18, 2012 and August 20, 2012 (three days overall during the electoral campaign, including Quebec's political party leaders' debate which took place on August 19, 2012). All the extracted tweets are in French. We then manually created a list of authoritative users whose importance on the selected topic (that is, 2012 Quebec election) is relevant. We have identified 76 authoritative users. These users correspond to active official representatives of political parties on Twitter, influential political analysts and well-known Quebec personalities, who are active on Twitter, from various fields (e.g. art, business, government). In our investigation, we found that the contents posted by these users is widespread and their tweets were engaged toward a point of view. Next, based on the collected tweets, we manually labeled 828 users as non-authoritative by examining their Twitter profiles and reading several of their posts. As a result of this process, our data set contains 904 users; 76 users (8.4% of the whole data set) among them were labeled as authoritative and 828 users were labelled as non-authoritative. Finally, it is important to note that, since the classification labeling process relies on human judgment, which implies examining hundreds of user profiles and reading numerous tweets, we had to set a limit on the number of users in our labeled collection.

For users features, we considered the following four metrics:

(1) The number of followers of a user, which indicates the size of the audience for that user [Cha et al. 2010]. Users having a large number of followers are, in general, popular users who act as a successful broadcast medium since their tweets are read by every follower.
(2) The Followers to Followees ratio (F-F ratio), that is, the number of a user's followers and the number of other people that the user follows (followees). As suggested in [Bigonha et al. 2012], [Leavitt et al. 2009], this metric valorises users who are widely followed, but have some selection for following others. In fact, the F-F ratio approaches infinity when the number of followers is very high and the number of followees is very low. This is the case for authoritative users who are followed by a large number of tweeters but are selective in following others. On the other hand, when the ratio approaches 0, the user might be categorised as a conversationalist, one who follows more users than is followed [Bigonha et al. 2012], [Leavitt et al. 2009].
(3) The number of retweets, which measures the number of times an author's tweets were retweeted by other users. This metric indicates the ability of a user to generate content with pass-long value [Cha et al. 2010]. The action of retweeting is thus an evidence for influence that has occurred, since a user is influenced to reproduce the content [Pal and Counts 2011]. Authoritative users tend to have a large number of their posts retweeted by others.
(4) The number of mentions, which is measured by the number of times a user was cited or had her tweet replied to. This metric indicates the ability of a user to engage others in a conversation [Cha et al. 2010] and also reflects the impact of that user in the system with respect to the topic of interest. Mentioning could be thus considered as a mark of authority. Authoritative users are those who have been most frequently mentioned.

Experiment 1. Similar to the experiments on Stack Exchange data sets, the aim of this first set of experiments is to evaluate our approach using different subsets of the Twitter features (number of followers, F-F ratio, number of retweets and number of mentions) considered in this paper. In our experiments, we constructed several data sets using the following subsets of features: (1) number of followers and F-F ratio, (2) number of followers and number of retweets, (3) number of followers and number of mentions, (4) number of retweets and number of mentions, (5) number of followers, number of retweets and number of mentions, (6) number of followers, F-F ratio and



(a) Number of followers and F-F ratio.

(b) Number of followers and number of retweets.

(c) Number of followers and number of mentions.

(d) Number of retweets and number of mentions.

Fig. 5: Density curves of several $2D$ user features combinations over Quebec Election data.

Input features	Accuracy	CD	FA	F-measure
# of followers and F-F ratio	95.2%	97.3%	4.9%	0.774
# of followers and # of retweets	97.3%	98.6%	2.7%	0.862
# of followers and # of mentions	98.5%	100%	1.5%	0.921
# of retweets and # of mentions	97.3%	94.7%	2.4%	0.857
# of followers, # of retweets and # of mentions	97.3%	97.3%	2.6%	0.860
# of followers, F-F ratio and # of retweets	97.4%	98.6%	2.6%	0.867
F-F ratio, # of retweets and # of mentions	97.3%	98.6%	2.7%	0.862
All features	**99.2%**	**97.3%**	**0.6%**	**0.954**

Fig. 6: Performance results over Quebec Election data.

number of retweets, (7) number of followers, F-F ratio and number of mentions, and, finally, (8) the four user features. Then, for each constructed data set, we used our approach to identify authoritative users. To this end, we set C_max to 6 in all our experiments and selected the optimal number of components that minimize ICL-BIC. Interestingly, as with the experiments on Stack Exchange data sets, we found that the number of components varies from three to four. For the purpose of illustration, Figure 5 shows the estimated probability density function of the users' vectors resulting from several $2D$ features combinations. The multivariate beta component that represents the highest feature values corresponds to authoritative users.

In Figure 6, we illustrate the performance results over the Quebec Election data, based on several features combinations. Shaded regions in this figure correspond to the best values of the four evaluation metrics considered in this paper. As can be seen from this figure, the combination of the four Twitter user features provides the best Accuracy, FA rate and F-measure. Interestingly, in comparison to the results provided by other feature subsets, the combination of the four features contributes to lowering the FA rate (0.6%) and maintains a high CD rate (97.3%). The highest CD rate

Algorithm	Accuracy	CD	FA	F-measure
Proposed	**99.2%**	**97.3%**	**0.6%**	**0.954**
AdaBoost	99.2%	98.7%	0.7%	0.955
Bagging	98.8%	94.7%	0.7%	0.935
Decorate	99.4%	97.4%	0.4%	0.967
LogitBoost	98.8%	94.7%	0.7%	0.935
MultiBosstAB	98.6%	94.7%	1%	0.923
ADTree	99%	96.1%	0.7%	0.942
Random Forest	99.3%	98.7%	0.6%	0.962
RBF Network	97.7%	92.1%	1.7%	0.875

Fig. 7: Accuracies of compared algorithms on Quebec Election data.

value (100%) is achieved when combining the number of followers and the number of mentions. However, using these latest two features, a non-negligible fraction of non-authoritative users (FA rate = 1.5%) is misclassified as authoritative. Overall, this first experiment suggests that the combination of the number of followers, F-F ratio, number of retweets and number of mentions provide the best trade-off between CD rate and FA rate and achieves the highest accuracy and F-measure.

Experiment 2. Our goal now is to compare the performance of our approach to that of AdaBoost, Bagging, Decorate, LogitBoost, MultiBosstAB, ADTree, Random Forest and RBF Network. We did not consider SVM in this experiment since the algorithm classifies all users as non-authoritative and was not able to identify any authoritative users. Note that, for all compared algorithms, we present results using the four Twitter user features considered in this paper (that is, number of followers, F-F ratio, number of retweets, number of mentions). Figure 7 summarizes the results of the comparison. Shaded regions correspond to the best Accuracy, CD rate, FA rate and F-measure values. As can be seen from this figure, except for RBF Network, competing algorithms provide comparable and accurate results. The best accuracy (99.4%), FA rate (0.4%) and F-measure (0.967) were achieved by Decorate. On the other hand, the highest CD rate (98.7%) is achieved by AdaBoost and Random Forest. As depicted by Figure 7, such results are very close to those provided by our approach. Finally, from Figure 7, we note that the performance of RBF Network, is less competitive compared to the other algorithms. However, even though this algorithm achieves the lowest Accuracy, CD rate and F-measure values and the highest FA rate, we believe that it's performance is acceptable.

To summarize, the experiments conducted on data sets extracted from Stack Exchange and Twitter suggest that all competing algorithms provide meaningful results. A general perception in the field seems to be that a supervised method works better than an unsupervised one, since it is able to exploit information about the grand truth provided by humans, which is not available to an unsupervised approach. The experiments presented in this section show that our unsupervised method performs as well as (and sometimes better than) several supervised approaches. But our approach also has the ability to mine unlabeled data, a considerable practical advantage for real-world applications in which class labels are not available.

5. CONCLUSION

In this paper, we have discussed some drawbacks of existing authoritative user identification approaches including their incapability to automatically discriminate between authoritative and non-authoritative users, their dependency on labeled data, and their need for user parameters which are difficult to tune. To address these problems, we have proposed a mixture model-based approach to automatically identify authoritative users. In our approach, we first propose to represent each user as a feature vector such that each element of the vector contains information that would reflect the au-

thority of a user in an online community. Next, we model these vectors as a mixture of multivariate beta distributions. The number of components is estimated using the integrated classification likelihood Bayesian information criterion, while the parameters of the mixture are estimated using the EM algorithm. Such an approach allows the identification of the multivariate beta component containing the most authoritative users. We evaluated the suitability of our approach in tests and comparisons with some supervised methods, using real data extracted from the Stack Exchange question-answering network and from Twitter. The results showed that the proposed approach yields high-quality results. As a matter of fact, our unsupervised authoritative users identification method exhibits results that are comparable (and, in several cases, even superior) to those of supervised attribute learning algorithms.

Finally, it is worth noting that, in contrast to most existing authoritative users detection methods, our approach has several practical advantages. As discussed earlier, the proposed method is parameterless which is, in turn, a considerable advantage in practice. Parameter-laden methods are, however, critical and their application to real situations is not obvious since it is rarely possible for users to supply the parameter values accurately. Furthermore, the method presented in this paper does not require labeled samples or prior knowledge about the data under investigation to detect authorities. In fact, our approach is able to automatically identify authoritative from non-authoritative users while many existing approaches provide a ranked list of users only without formally specifying how many users should be chosen as authoritative from the ranked list. We believe that these notable features of the proposed approach provide significant evidence about its practicality. The experiments conducted on different real data sets corroborate our claim.

ACKNOWLEDGMENTS

The authors would like to thank Eric Beaudry, Jean-Eudes Blin and Morgan Steunou for gathering and preparing the data sets used in the experiments. The authors also would like to thank the reviewers and associate editor for their valuable comments and important suggestions.

REFERENCES

N. Agarwal, H. Liu, L. Tang, and P. S. Yu. 2012. Modeling Blogger Influence in a Community. *Social Network Analysis and Mining* 2, 2 (2012), 139–162.

L.J. Bain and M. Engelhardt. 2000. *Introduction to Probability and Mathematical Statistics* (second ed.). Duxbury Press.

J.C. Bezdek. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.

C. Bigonha, T. N. C. Cardoso, M. M. Moro, M. A. Gonalves, and V. A. F. Almeida. 2012. Sentiment-Based Influence Detection on Twitter. *Journal of the Brazilian Computer Society* 18, 3 (2012), 169–183.

F. Bonchi, C. Castillo, A. Gionis, and A. Jaimes. 2011. Social Network Analysis and Mining for Business Applications. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (2011).

M. Bouguessa. 2011. An Unsupervised Approach for Identifying Spammers in Social Networks. *Proc. 23rd IEEE Int'l Conf. Tools with Artificial Intelligence (ICTAI'2011)* (2011), 832–840.

M. Bouguessa, B. Dumoulin, and S. Wang. 2008. Identifying Authoritative Actors in Question-Answering Forums: The Case of Yahoo! Answers. *Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD'2008)* (2008), 866–874.

M. Bouguessa, S. Wang, and B. Dumoulin. 2010. Discovering Knowledge-Sharing Communities in Question-Answering Forums. *ACM Transactions on Knowledge Discovery from Data* 5, 1 (2010).

N. Bouguila, D. Ziou, and E. Monga. 2006. Practical Bayesian Estimation of a Finite Beta Mixture Through Gibbs Sampling and its Applications. *Statistics and Computing* 16, 2 (2006), 215–225.

S. Boutemedjet, D. Ziou, and N. Bouguila. 2011. Model-Based Subspace Clustering of Non-Gaussian Data. *Neurocomputing* 73, 10-12 (2011), 1730–1739.

S. Budalakoti and R. Bekkerman. 2012. Bimodal Invitation-Navigation Fair Bets Model for Authority Identification in a Social Network. *Proc. 21st ACM Int'l Conf. World Wide Web (WWW'2012)* (2012), 709–718.

M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. *Proc. AAAI Int'l Conf. Weblogs and Social Media (ICWSM'2010)* (2010).

Z. Cheng, J. Caverlee, and K. Lee. 2013. A Content-Driven Framework for Geolocating Microblog Users. *ACM Transactions on Intelligent Systems and Technology* 4, 1 (2013).

A. Dempster, N. Laird, and D. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society, (Series B)* 39 (1977), 1–37.

M.A.T. Figueiredo and A.K. Jain. 2002. Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 3 (2002), 381–396.

S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. P. Gummadi. 2012. Cognos: Crowdsourcing Search for Topic Experts in Microblogs. (2012).

Y. Ji, C. Wu, P. Liu, J. Wang, and K.R. Coombes. 2005. Applications of Beta-Mixture Models in Bioinformatics. *Bioinformatics* 21, 9 (2005), 2118–2122.

W.-C. Kao, D.-R. Liu, and S.-W. Wang. 2010. Expert Finding in Question-Answering Websites: A Novel Hybrid Approach. *Proc. ACM Symposium on Applied Computing (SAC'2010)* (2010), 867–871.

S. Kumar, F. Morstatter, and H. Liu. 2013. *Twitter Data Analytics*. Springer, New York, NY, USA.

H. Kwak, C. Lee, H. Park, and S. Moon. 2010. What is Twitter, a Social Network or a News Media? *Proc. 19th ACM Int'l Conf. World Wide Web (WWW'2010)* (2010), 591–600.

A. Leavitt, E. Burchard, D. Fisher, and S. Gilbert. 2009. The Influentials: New Approaches for Analyzing Influence on Twitter. *a publication of the Web Ecology Project* (2009).

J. Liu, Y.-I. Song, and C.-Y. Lin. 2011. Competition-Based User Expertise Score Estimation. *Proc. ACM SIGIR Int'l Conf. Research and Development in Information Retrieval (SIGIR'2011)* (2011), 425–434.

Z. Ma. 2011. *Non-Gaussian Statistical Models and Their Applications*. Ph.D. Dissertation. KTH - Royal Institute of Technology, Stockholm.

Z. Ma and A. Leijon. 2009. Beta Mixture Models and the Application to Image Classification. *Proc. 16th IEEE Int'l Conf. Image Processing (ICIP'2009)* (2009), 2045–2048.

I. Olkin and H. Rubin. 1964. Multivariate Beta Distributions and Independence Properties of the Wishart Distribution. *The Annals of Mathematical Statistics* 35, 1 (1964), 261–269.

A. Pal. 2012. *User Classification in Online Community*. Ph.D. Dissertation. University of Minnesota.

A. Pal and S. Counts. 2011. Identifying Topical Authorities in Microblogs. *Proc. 4th ACM Int'l Conf. Web Search and Data Mining (WSDM'2011)* (2011), 45–54.

A. Pal, R. Farzan, J. A. Konstan, and R. E. Kraut. 2011. Early Detection of Potential Experts in Question Answering Communities. *Proc. 19th Int'l Conf. User Modeling, Adaption and Personalization (UMAP'2011) - Lecture Notes in Computer Science* 6787 (2011), 231–242.

A. Pal, F. M. Harper, and J. A. Konstan. 2012. Exploring Question Selection Bias to Identify Experts and Potential Experts in Community Question Answering. *ACM Transactions on Information Systems* 30, 2 (2012).

F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios. 2012. Influence and Passivity in Social Media. *Proc. 21st ACM Int'l Conf. World Wide Web - Companion Volume (WWW'2012 Companion)* (2012), 791–798.

D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. 2011. Influence and Passivity in Social Media. *Proc. 20th ACM Int'l Conf. World Wide Web - Companion Volume (WWW'2011 Companion)* (2011), 113–114.

X. Tang and C. C. Yang. 2012. Ranking User Influence in Healthcare Social Media. *ACM Transactions on Intelligent Systems and Technology* 3, 4 (2012).

J. Weng, E.-P. Lim, J. Jiang, and Q. He. 2010. TwitterRank: Finding Topic-sensitive Influential Twitterers. *Proc. 3rd ACM Int'l Conf. Web Search and Data Mining (WSDM'2010)* (2010), 261–270.

C.-L. Yang and Y.-H. Chen-Burger. 2012. On-Line communities making scense: a hybrid micro-blogging platform community analysis framework. *Proc. 6th KES international conference on Agent and Multi-Agent Systems: technologies and applications (AMSTA'2012)* (2012), 134–143.

J. Zhang, M. S. Ackerman, and L. Adamic. 2007. Expertise Networks in Online Communities: Structure and Algorithms. *Proc. 16th ACM Int'l Conf. World Wide Web (WWW'2007)* (2007), 221–230.

Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao. 2009. Routing Questions to the Right Users in Online Communities. *Proc. IEEE Int'l Conf. Data Engineering (ICDE'2009)* (2009), 700–711.

H. Zhu, H. Cao, H. Xiong, E. Chen, and J. Tian. 2011. Towards Expert Finding by Leveraging Relevant Categories in Authority Ranking. *Proc. 20th ACM Int'l Conf. Information and Knowledge Management (CIKM'2011)* (2011), 2221–2224.