Finding Influential Users in Twitter Using Cluster-Based Fusion Methods of Result Lists

Alexandros Georgiou, Andreas Kanavos, Christos Makris

Finding Influential Users in Twitter Using Cluster-Based Fusion Methods of Result Lists

Alexandros Georgiou, Andreas Kanavos, and Christos Makris

Computer Engineering and Informatics Department
University of Patras, Patras, Greece
{georgiua,kanavos,makri}@ceid.upatras.gr

Abstract. The topic of the paper is to present a novel methodology in order to characterize influential users, such as members of Twitter, as they arise in social networks. The novelty of our approach lies in the fact that we incorporate a set of features for characterizing social media authors, including both nodal and topical metrics, along with new features concerning temporal aspects of user participation on the topic. We also take advantage of cluster-based fusion techniques for retrieved result lists for the ranking of top influential users.

Keywords: Cluster-based Methods, Influential Users, List Fusion Methods, Social Networks, Temporal Features, Twitter, Web Mining

1 Introduction

The task of finding the most influential users in an online social networking environment has gained a great amount of attention in recent years. Special focus is given on social networking platforms called microblogging platforms. These platforms allow only short messages to be published (usually ranging in a few hundred characters), a fact that raises a wide range of problems against text-based information retrieval techniques.

A prominent example of such microblogging platforms is the Twitter online social network which only allows messages of 140 characters maximum. Twitter is an internationally famous social networking platform with hundreds of millions of active users. Each user can create an unlimited circle of affiliated users to whom they can publish updates (called *tweets*). Users are additionally presented with a list of tweets by their affiliated users sorted by the latest, called *timeline*. User relations in Twitter are not necessarily reciprocal: user a may *follow* user b, without user b having to authorize it or to follow back. When user b chooses to follow back user a, users a and b can be called *friends*. The Twitter platform allows users to repost content that they find interesting, an action called *retweet* which is signified by the characters "RT" following the original content producer's username. A user is able to directly mention another user with the character "@" followed by the mentioned user's username. Topics of discussion can be initiated by any user and organized around user-specified keywords, called *hashtags* and signified by the character "#" followed by the desired keyword.

Recent studies [10], [23] have shown that groups of intermediate level users act as propagating nodes for the information flow on such networks, and users rely preferably

on other users or special purpose user accounts for their information about certain topics. Taking into account the spread of such online social networks and the impact that they have on many aspects of everyday social, economic and political reality, identifying users with high influence around specified topics is of crucial importance for social media marketing agents, governments, policy makers, celebrities and communities.

The rest of the paper is structured as follows. Section 2 presents background topics while Section 3 presents our methodology followed and the system developed. In Section 4, details of the implementation of the system as well as the evaluation study conducted and the results gathered on both the sentiment analysis topic and the community detection topic are presented. Finally, Section 5 concludes our work and presents directions for future research.

2 Related Work

Recently, the identification of topical (or influential) authorities in microblogging has gained a lot of attention. In [19], the challenge of finding the most interesting and authoritative authors for any given topic in Twitter is reported. Authors provide a set of features for characterizing any social media author, including both nodal and topical metrics. Their experimental results show that a probabilistic clustering over a feature space, followed by a within-cluster ranking procedure, can yield to a final list of top authors for a given topic. More specifically, their technique uses a Gaussian Mixture Model to cluster users into two clusters over their feature space as the aim is to reduce the size of the target cluster; that is the cluster containing the most authoritative users. In addition in [11] and [12], the notion of influence from users to networks is extended and in following, personality as a key characteristic for identifying influential networks is considered. The system creates influential communities in a Twitter network graph by considering user personalities where an existing modularity-based community detection algorithm is used. At a later point, the insertion of a pre-processing step that eliminates graph edges based on user personality is utilized. Moreover in [13], an efficient and innovative methodology for community detection that will also leverage users' behavior on emotional level is introduced.

Interesting is the work presented in [22], which employs Latent Dirichlet Allocation and a variant of the PageRank algorithm that clusters according to topics and finds the authorities of each topic; the proposed metric is called TwitterRank. The field of analysis in social networks is related to link analysis in the web with cornerstone the analysis of the significance of web pages in Google using the PageRank citation metric [18], the HITS algorithm proposed by Kleinberg [15] as well as their numerous variants discussed in [16]. PageRank employs a simple metric based on the importance of the incoming links while HITS uses two metrics emphasizing the dual role of a web page as a hub and as an authority for information.

Historically, the above as well as other approaches and techniques have been harnessed throughout microblogging areas. In [8], an overall generative model for questions and answers in community-based Question Answering (cQA) services is developed, which is then altered to obtain a novel computationally tractable Bayesian network model. Initially, they seek to discover latent topics in the content of questions as

well as the associated answers, and latent topic interests of users. Then, they recommend answer providers for new questions according to discovered topics as well as term-level information of queries and users. What is more, in [17], authors present an investigation dealing with user perceptions about credibility tweets, where they examined key elements of the information interface for their impact on credibility judgements. Their results indicate that users had difficulty determining the truthfulness of content and that their judgement was clouded and often based on heuristics (e.g. if a post has been retweeted) and biased systematically (e.g. topically-related user names seen as more credible).

Furthermore, the similar problem though in other platform (e.g. in in Yahoo! Answers) was addressed in [5]. Their method automatically discriminates between authoritative and non-authoritative users through modeling the authority scores of users as a mixture of gamma distributions. The number of components in the mixture is estimated by the Bayesian Information Criterion (BIC) while the parameters of each component are estimated using the Expectation-Maximization (EM) algorithm. Concerning Yahoo! Answers, authors in [2] investigated methods for exploiting specific community feedback so as to automatically identify high quality content. More in detail, a general classification framework for combining the evidence from different sources of information, that can be tuned automatically for a given social media type and quality definition, is proposed and the experiments show an accurate separation of high-quality items from the rest, non-notable.

Finally, relative study with the current one is [3] by Anderson et al. in which it is investigated whether similarity in the characteristics of two users can affect the evaluation that one user provides to another. They analyze this problem under a range of natural similarity measures, demonstrating how the interaction between likeness and status can produce strong effects. Among these measures is a resemblance of interests using a distance metric capturing overlap in the types of content that users produce, as well as a similarity of social ties using a measure of the overlap in the sets of people they evaluated.

3 System Description

3.1 Modular Architecture

In the social media mining system we developed, the most authoritative users per topic are identified based on a variety of features that combine the quality of content they provide. Text similarity measures, social impact through retweets, ability to spike conversations considering the content provided (through conversational tweets), social graph relations and time-related variables measuring frequency and timezone span consist important characteristics as well.

Our system architecture consists of the following modules:

- *A Twitter access module*: Twitter database is accessed through Twitter API by this module, using the Twitter4j Java library for Twitter application development. This module receives topic name (#hashtag) as input, and returns user tweets from the specific topic as well as active user data and social graph relations from the total Twitter social graph.

- *A Parser module*: Output from the Twitter access module is parsed to create appropriate username searchable hashmaps which include all tweets, social graph data and time-related data. This stage is necessary as a preparation for the feature extraction process.
- *Feature extraction module*: Hashmaps containing username - tweet set pairs are given as input from the Parser module. Numbers of original tweets, retweets, conversational tweets are counted, social graph relations are measured, posting frequency for each user is reported and tweets are distributed into four 8-hour time zones (morning, noon, evening, night) based on standard Twitter timestamps. These counts and measures are later combined to create the list of features for every user who participates in the specific topic. Hashmaps are restructured to contain username - feature value pairs.
- *A Clustering module*: the set of username - feature values hashmaps is given as input in a module responsible for the clustering algorithms. Using Fuzzy C-Means, data clusters are created.
- *A Ranking module*: Different types of ranking techniques are compared at the clustered user data. Gaussian ranking used by [19] is tested against a method described in [14].

This system operates nearly *on the fly*, in the sense that database read-write operations are used only for back-tracking reasons and result storage. Since the data size of specific topics is average and Twitter outputs its content in JSON form, an average computer system is able to execute hashmap counts and feature extraction in memory. There is an open window for parallelization at this point, discussed in Section 5. Direct access to the Twitter dataset queried by topic was used, through the requests documented in the Twitter API. Topic is user-defined at the beginning of the execution, but the Twitter API presents limitations on the maximum data transactions per hour.

3.2 Feature Extraction

This subsection describes the set of features we inherited from [19] (named "Basic Features") and our contribution to the feature set, which is named "Time-based Features".

3.2.1 Basic Features User features are extracted by calculating and combining different measures, as proposed in [19]. Thus, we get measures of Original Tweets, meaning new content provided by the user, Conversational Tweets, meaning replies to user (signified by the "@username" string), Repeated Tweets, meaning content that the user provided and is then reproduced by other users (signified by the "RT" string), Mentions, meaning unique references to user's username by other users and Graph Characteristics, meaning measures of total and topic-active friends and followers of the specific user.

According to this method, for the given topic we calculate the following features:

- *Topical signal (TS)* indicates the percentage of participation in a given topic by a specific author, regardless of the type of tweets.
- *Signal strength (SS)* shows how strong an author's topical signal is based on how many tweets of this author have original content.

- *Non-Chat signal (∼CS)* tries to capture how many of the author's tweets are not involved in a direct conversation with friends or followers. This is used to discard any conversations that the specific author participated in but were not initiated by them. This feature involves an λ parameter calculated approximately at 0.05 to satisfy the constraint mentioned above.
- *Retweet impact (RI)* demonstrates the impact of content generated by the author under measurement. The number of retweets is considered directly proportional to the impact this content has over the community around the specific topic. The calculations use multiplication by a logarithmic function to rule out the impact that may be generated by overly supportive followers of the specific author.
- *Mention impact (MI)* is counting how much an author is mentioned during the discussion of a certain topic, indicating that they are socially regarded as an authority in the topic. A log function is included here too, to ensure that the author is not mentioned due to their mentioning other authors (in a conversational manner).
- *Information Diffusion (ID)* is a social graph - based feature showing the ratio of number of users activated by the author on log-scale. We consider that an author is "activated" if they start tweeting on a topic after another user from the user's network that has tweeted on the topic before the author.
- *Network score (NS)* is a mere social graph - based feature which counts the number of users active on the topic that are in the social circle of the author.

For further details on the measurements and the calculations involved in the basic feature set, one should refer to [19].

3.2.2 Time-based Features A central point of motivation for this paper is that the dimension of time is absent from any measure extracted from Twitter topics. This type of topic analysis is based on a static idea about the topic data: it takes topic discussions as solid data, showing indifference for temporal distribution, namely the way that discussion data is spread through time. The reality of social media topic discussion is more dynamic than this. Sparks of "discussion traffic" can be recognized when the topic is "hot" meaning that at some time intervals, due to events of conjuncture, a lot of users get attracted by the specific topic. This can lead some users getting "authoritativeness" points for a short period of activity in the topic's lifecycle. Our claim is that a strongly authoritative user should provide content or be conversationally active throughout the total lifecycle of a topic. In addition, authoritative user tweets should be discoverable throughout the day, so that users active in different time zones could interact with the authoritative user content. This is true especially for topics with a lifecycle that lasts days or months and for topics that have global interest attracted to them, such as an economic or political crisis topic, sports organization topics, etc.

We consider zero time according to the timestamp of the first tweet containing the requested #hashtag and ending time according to the timestamp of the last such tweet by the time of query. We propose new features that put into consideration the above mentioned parameters:

- *Frequency* is a feature indicating the contribution of a specific author in a topic during the entire lifecycle of the topic. In our approach, high values of tweeting

frequency increase the authority of the author. This may seem contrary to the burst of information in short time segments that usually emerge in social networks, but we claim that for a user to be more authoritative, their content generation must follow and span a large percentage of the topic lifecycle. In the example which motivates the research in [19], the *Gulf of Mexico Oil Spill*, Twitter accounts of environmental agencies considered authoritative for this topic should keep their followers informed as long as the topic is active. High frequency scores can rule out effects of posting burst. To calculate posting frequency, the ratio

$$freq = \frac{tweets_i}{endtime_{topic} - starttime_{topic}} \tag{1}$$

is used for every author active in the topic.

– *Part-of-day measure* captures the notion of users participating in a discussion from different time zones. This is especially interesting for topics with global effect and global audience. Due to the design of a platform such as Twitter, when a user logs in the platform, they see content in a newer-to-older fashion. To discover older content they have to scroll down, even if a search-by-topic approach is utilized. If time zones are taken into account, a user in East Asia should scroll down a lot to read original content from an author posting from the United States (taking into account that most users are not $24/7$ online). For a global notion of authority, an author (such as an account registered by an institution or a news agency) should have a posting distribution that covers all day. This is an approximation feature; therefore dividing in four 6-hour parts - of - day measures (morning, noon, evening, night) is enough to demonstrate such distribution. In each part - of - day, simple count of tweets is used and provided as a clustering dimension.

3.3 Clustering and Ranking

For the clustering and ranking process, used to derive possible authorative users, two methods were compared: *(i)* clustering and ranking with the use of Gaussian Mixture Models (GMM) and the Expectation - Maximization (EM) algorithm (the method used in [19], and *(ii)* our proposal, clustering and ranking with the use of cluster-based fusion of retrieved lists (as presented in [14]. Our proposal also contains the substitution of the simple K-means algorithm for primal clustering by the Fuzzy C-means (as found in [4], [7]) algorithms because of the notion of similarity it points out which is well suited when one has to deal with user content on a specific topic.

3.3.1 Gaussian Mixture Model A Gaussian Mixture Model (GMM) is a probability density function calculated as the weighted sum of Gaussian component densities. More specifically, a GMM is a weighted sum of M component Gaussian densities as given by the equation,

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \sum i) \tag{2}$$

where x is a D-dimensional data vector of features, wi are the mixture weights and $g(x|\mu_i, \sum i)$ are the component Gaussian densities. Each is a D-variate Gaussian function with mean vector and covariance matrix.

GMMs are mostly used in continuous-value contexts, i.e. speaker recognition systems and biometric data. This raises a conceptual issue concerning the use of a GMM in the aspect of ranking authors in a microblogging environment. It is not proved that the set of features discussed in the previous section follows the normal (or Gaussian) distribution. Intuition and experiments show that a small cluster of authors around a specific topic achieves great scores, while a long tail of authors achieve low scores. Normal distribution implies that most of the authors should be at a $+-s$ distance from the average score (where s is standard deviation), which is not the case especially for popular topics with thousands of followers. Most of the followers participate through a low activity of retweets or commentary tweets, while authorative users should have frequent multi-type contribution on the topic.

3.3.2 Using cluster-based fusion of retrieved lists The technique of cluster-based fusion is presented and evaluated in [14]. The key concept of this technique is that inter-similarity of documents presented in different query result lists should be rewarded. Given a query q, a document d and a corpus of documents C, one can get $L_1, \ldots, L_m$ result lists on m retrievals based on query q. In these lists, d may appear in a low position in a result list. Straightforward list fusion methods, such as the CombSum, CombMNZ and Borda methods use partial list rankings to build a final result list, which can lead to very low total ranking [14], of an important document d. Cluster-based fusion uses the *cluster hypothesis* to reward low-ranked documents with the condition that they belong in the same cluster with high ranked documents. Therefore, the cluster-based fusion method runs *some* clustering algorithm on the document set of documents appearing in the partial list and calculates the final ranking list based on partial list score plus cluster score.

In our proposal, we utilize this method using the fuzzy C-means algorithm for clustering documents. More specifically, the results are initially clustered into k lists using the fuzzy C-means algorithm, which permits an author to appear in more than one list. Each list is sorted with the Gaussian ranking method and then the cluster-based fusion method calculates the fusion score of the final ranking list. The cluster-based fusion method in our setting runs for the ClustFuseCombSUM, ClustFuseCombMNZ and ClustFuseBorda [14] best-performing versions of the algorithm.

4 Experimental Evaluation

In the next three subsections, the experimental setting for our approach is presented (subsection 4.1), followed by the results for the top-10 influential users of different versions of the algorithm (subsection 4.2) and results of anonymous user evaluation (subsection 4.3). The logic behind the experiments is to evaluate the quality of results between the GMM-based approach and the cluster-based fusion approach (with different versions of fusion strategies).

4.1 Dataset

For the construction of our test data set, we had to respect the current limitations of the Twitter API, together with the need to build a data set of topics that have differences in their temporal development. The Twitter database was queried for the hashtags: #blacklivesmatter, #bigdata and #germanwings.

The first hashtag, #blacklivesmatter, responds to a discussion topic about a social situation with duration in time and very different activity levels from time to time. The second hashtag, #bigdata, is reflecting a discussion topic with mostly scientific and business interest and quite sparse but also quite linear activity in time. The third hashtag, #germanwings had to do with an emerging tragic event and organized a discussion topic that demonstrated a burst of activity for the first few days but then faded to very low activity levels.

The construction of the data set was completed with a two-step repetitive process where firstly a tweet was returned as answer to the hashtag query and then a second query was performed to get the friends and followers list of the user that posted the tweet. That process resulted in 2.000 tweets and 49 user accounts (with a total of 50.622 followers) for the topic #blacklivesmatter, 2.000 tweets and 45 user accounts (with a total of 98349 followers) for the topic #bigdata and 1.860 tweets and 40 user accounts (with a total of 86.002 followers) for the topic #germanwings.

4.2 Top-k Users

For each topic and each tweet on the data set, two sets of experiments were conducted. The first set of experiments produced top-k ranked user lists by the execution of the GMM-based version of the algorithm as presented in [18] and three versions of cluster-based fusion algorithms using the ClustFuseCombMNZ, ClustFuseBorda and ClustFuseCombSUM strategies for list fusion, as presented in [14], without the addition of the proposed temporal features. In the four columns of Table 1, one can see the top-5 ranked user lists for the three different topics. The second set of experiments produced top-10 ranked user lists like the first set, but this time including the temporal features we proposed in section 3.2.2. The four columns of Table 2 present the results of the four different algorithms for the top-5 ranked user accounts. It is important to note here that there are differences in the ranking produced by the algorithms after the addition of the temporal features, mostly affecting the methods based on the cluster hypothesis (e.g. ClustFuseCombSUM).

On the other hand, as previously mentioned, the average number of Followers per Community is slightly lower when the emotional methodology is followed. This is mainly a result of the way that Influential Metric is defined as it deals with an overall estimation of the impact of each user in the produced community.

4.3 User Evaluation

For the purposes of user evaluation of the different result sets, we organized an online survey and asked social media users to anonymously complete some web forms. A special occasion web application was developed linked to a database where answers where

Table 1. Top-5 ranked users with temporal features

GMM	ClustFuseCombMNZ	ClustFuseBorda	ClustFuseCombSUM
#blacklivesmatter			
Shgamha	_PoeticRebel	Me_MrCool	Shelby_ville
newBREED_	pces	foodbruh_	chilllaxx_
ArtisMentis	I_Cant_Breathe	Shelby_ville	dmwwalker343
_PoeticRebel	Shgamha	chilllaxx_	AshhhG_
I_Cant_Breathe	newBREED_	dmwwalker343	newBREED_
#bigdata			
AnRcloudSoft	PyramidAnalytic	eberman007	revistadircom
revistadircom	bobehayes	GammaAnalytics	phatpenguin
danablouin	ThugMetricsNews	ThugMetricsNews	byod_news
METAMORF_US	aleson_es	KobbyDon1	BusinessNWSRM
phatpenguin	ymtreb	mallys_	BDUGUK
#germanwings			
GAABY	GAABY	DobleYouu	DobleYouu
WSJIndonesia	WSJIndonesia	FresaaChampagne	FresaaChampagne
KeystoneIDEAS	die_politik	EkoPardiyanto	EkoPardiyanto
mycomfor	mycomfor	adrianaeloca	adrianaeloca
EkoPardiyanto	lesatorr	nonotina	nonotina

concentrated for later process. The evaluation scenario complied with the following assumptions: (1) evaluating users were anonymous (age and gender data where recorded for statistical reasons), (2) evaluating users were not presented with the results of the algorithms and are asked to rank usernames without guidance.

Users were presented with the whole data set and enabled to browse through the tweets, filter them by topic and query them by keyword or by username. After browsing through the data set, users were asked to choose the most influential username per topic, according to what they believe. That username was awarded by 10 extra points. After choosing the top username, users were presented with three forms, one for each topic, where they were asked to rank each of the usernames participating in the topics with a rank between 1 to 10 according to whether they are authoritative or not. The final rank for a username is the sum of ranks it has gained. A total number of 296 social media users from Facebook and Twitter took part in the evaluation survey with average age of 28.3 years and 37% of them were women. To understand the effectiveness of each method under evaluation, and also the effectiveness of the new time-based features we proposed, we used precision and Pearson - correlation metrics to measure the correctness of the algorithmic results and whether there is an agreement between method and user evaluation for the ranking order of users.

Precision and Pearson - correlation metrics are presented in Tables 3 and 4 for the two sets of experiments described in Subsection 4.2. As we can see in both situations, the cluster-based methods score better than the GMM-based algorithm. The GMM-based algorithm seems to outrun the cluster-based fusion method only when ClustFuseCombMNZ strategy is used for fusion. Please notice that abbreviations have been

Table 2. Top-5 ranked users without temporal features

GMM	ClustFuseCombMNZ	ClustFuseBorda	ClustFuseCombSUM
#blacklivesmatter			
Shgamha	pces	Me_MrCool	Shelby_ville
newBREED_	I_Cant_Breathe	foodbruh_	_PoeticRebel
ArtisMentis	_PoeticRebel	Shelby_ville	chilllaxx_
_PoeticRebel	Shgamha	_PoeticRebel	dmwwalker343
I_Cant_Breathe	newBREED_	chilllaxx_	AshhhG_
#bigdata			
AnRcloudSoft	NoSQLDigest	byod_news	NoSQLDigest
revistadircom	SocialNewsCorp	BusinessNWSRM	revistadircom
danablouin	KobbyDon1	BDUGUK	ThugMetricsNews
METAMORF_US	PyramidAnalytic	AnRcloudSoft	phatpenguin
phatpenguin	Paxata	eberman007	GammaAnalytics
#germanwings			
GAABY	flores_crespo	FresaaChampagne	FresaaChampagne
WSJIndonesia	tedmohs	lesatorr	lesatorr
KeystoneIDEAS	PhilDeCarolis	adrianaeloca	adrianaeloca
mycomfor	HInstMH	Peterotul97	Peterotul97
EkoPardiyanto	die_politik	HInstMH	HInstMH

used space wisely, i.e. #blac for #blacklivesmatter, #bigd for #bigdata and #germ for #germanwings.

Table 3. Precision and Pearson - correlation with temporal features

	GMM	ClustFuse CombMNZ	ClustFuse Borda	ClustFuse CombSUM
Precision				
#blac	0,7	0,6	0,85	0,8
#bigd	0,6	0,6	0,8	0,8
#germ	0,6	0,5	0,75	0,75
Pearson - correlation				
#blac	0,45	0,47	0,57	0,55
#bigd	0,49	0,47	0,62	0,64
#germ	0,51	0,48	0,55	0,59

In the case of adding temporal features, one can see a significant improvement in the precision of every method, and an average improvement in the Pearson - correlation. The algorithms based on the ClustFuseBorda and ClustFuseCombSUM strategy seem to perform better in terms of recommendation quality.

Table 4. Precision and Pearson - correlation without temporal features

	GMM	ClustFuse CombMNZ	ClustFuse Borda	ClustFuse CombSUM
#blac	0,7	0,5	**0,8**	**0,8**
#bigd	0,6	0,6	0,8	**0,85**
#germ	**0,7**	0,5	**0,7**	**0,7**
Pearson - correlation				
#blac	0,43	0,44	**0,58**	0,52
#bigd	0,46	0,42	0,57	**0,66**
#germ	0,51	0,48	0,55	**0,59**

5 Conclusions and Future Work

In this paper, a novel approach to the problem of the discovery of topical influential users in a microblogging environment was presented and evaluated. The important advances of this research are the suggestion of fuzzy clustering and cluster-based fusion of user lists, together with the addition of time-based features that improve the overall precision and correlation scores. The list fusion approach circumvents possible drawbacks that the GMM-based methods have in cases that user features do not follow a normal distribution, a situation most common in social network environments. There is an open question of parallelization of the methods presented in this paper for the creation of a nearly real time authority discovery system.

The aspects of time in web and social network mining tasks are rather newly introduced but can gain potential due to the dynamic nature of these networks. Recent work on personalized user profile recommendation [1] and on event discovery in Twitter [20], expand the aspect of temporal dynamics in such environments.

For the discovery of influential users to be more accurate, one must comprehend the properties of the microblogging network and the behavior of the users, such as understanding collaborative behavior [9], analyzing why a tweet is likely to be retweeted [21] and decoding the social mechanism that explains why users with many followers are not necessarily the most influential [6].

References

1. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In: Proceedings of the 3rd International Web Science Conference (WebSci). pp. 1–8 (2011)
2. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM). pp. 183–194 (2008)
3. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Effects of user similarity in social media. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM). pp. 703–712 (2012)
4. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers (1981)

5. Bouguessa, M., Dumoulin, B., Wang, S.: Identifying authoritative actors in question-answering forums: The case of yahoo! answers. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). pp. 866–874 (2008)
6. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: The million follower fallacy. In: Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM) (2010)
7. Dunn, J.C.: A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3, 32–57 (1974)
8. Guo, J., Xu, S., Bao, S., Yu, Y.: Tapping on the potential of q&a community by recommending answer providers. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM). pp. 921–930 (2008)
9. Honeycutt, C., Herring, S.C.: Beyond microblogging: Conversation and collaboration via twitter. 2009 42nd Hawaii International Conference on System Sciences (HICSS) pp. 1–10 (2009)
10. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday 14(1) (2008)
11. Kafeza, E., Kanavos, A., Makris, C., Chiu, D.: Identifying personality-based communities in social networks. In: Advances in Conceptual Modeling. pp. 7–13 (2014)
12. Kafeza, E., Kanavos, A., Makris, C., Vikatos, P.: T-pice: Twitter personality based influential communities extraction system. In: IEEE International Congress on Big Data. pp. 212–219 (2014)
13. Kanavos, A., Perikos, I., Hatzilygeroudis, I., Tsakalidis, A.: Emotional community detection in social networks. Computers & Electrical Engineering 65, 449–460 (2018)
14. Khudyak Kozorovitsky, A., Kurland, O.: Cluster-based fusion of retrieved lists. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). pp. 893–902 (2011)
15. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46(5), 604–632 (1999)
16. Langville, A.N., Meyer, C.D.: Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, Princeton, NJ, USA (2006)
17. Morris, M.R., Counts, S., Roseway, A., Hoff, A., Schwarz, J.: Tweeting is believing?: Understanding microblog credibility perceptions. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW). pp. 441–450 (2012)
18. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66 (1999)
19. Pal, A., Counts, S.: Identifying topical authorities in microblogs. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM). pp. 45–54 (2011)
20. Stilo, G., Velardi, P.: Time makes sense: Event discovery in twitter using temporal similarity. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). vol. 2, pp. 186–193 (2014)
21. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: Proceedings of the 2010 IEEE Second International Conference on Social Computing (SOCIALCOM). pp. 177–184 (2010)
22. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: Finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM). pp. 261–270 (2010)
23. Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J.: Who says what to whom on twitter. In: Proceedings of the 20th International Conference on World Wide Web (WWW). pp. 705–714 (2011)