International Journal on Artificial Intelligence Tools
Vol. 28, No. 8 (2019) 1960010 (26 pages)

DOI: 10.1142/S0218213019600108

Estimating Twitter Influential Users by Using Cluster-Based Fusion Methods*

Andreas Kanavos†, Alexandros Georgiou‡ and Christos Makris§
Computer Engineering and Informatics Department, University of Patras
Rio, Patras, Greece, 26504
†*kanavos@ceid.upatras.gr*
‡*georgiua@ceid.upatras.gr*
§*makri@ceid.upatras.gr*

Received 11 December 2018
Accepted 25 October 2019
Published 4 December 2019

A considerable part of social network analysis literature is dedicated to determining which individuals are to be considered as influential in particular social settings. Concretely, Social Influence can be described as the power or even the ability of a person to yet influence the thoughts as well as the actions of other users. So, User Influence stands as a value that depends on the interest of the followers of a concrete user (via retweets, replies, mentions, favorites, etc.). This paper focuses on identifying such phenomena on the Twitter graph and on presenting a novel methodology for characterizing Twitter Influential Users. The novelty of our approach lies in the fact that we have incorporated 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 have also implemented cluster-based fusion techniques in order to retrieve result lists for the ranking of top influential users. Hence, results show that the proposed implementations and methodology can assist in identifying influential users, that play a dominant role in information diffusion.

Keywords: Cluster-based methods; graph mining; knowledge extraction; list fusion methods; social media analytics; temporal features; user influence; web mining.

1. Introduction

Given the global spread of social media usage over the last years, it has become clear that microblogging platforms, like Twitter, are being used as a means to either express or be informed about the personal views of many users on various subjects. Additionally, the simplification of posting and sharing content on these platforms

*A preliminary version of this paper was presented in 14th International Conference on Artificial Intelligence Applications and Innovations, AIAI 2018, Rhodes, Greece, May 25–27, 2018.

has led to the creation of vast amounts of information on a daily basis, so that the following question arise: "Is it possible to analyze all this content in such a way that certain of its aspects, like sentiment or numerical values, can be obtained in the form of meta-data? Can these meta-data be processed and used in such a way that future information from the exact source will be affected in a desired way?".

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 spreading in a few hundred characters), a fact that raises a wide range of problems against text-based information retrieval techniques, e.g. text mining, document clustering and ranking.

A prominent example of such microblogging platforms is the Twitter online social network which only allows messages of 140 characters maximum (with the latest expansion being up to 280 characters). 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. More to the point, "friends" is a list of users that user a *follows*, while "followers" is a list of users that *follow* user a.

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[15,36] 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. A survey is introduced in Ref. 31 where authors focus on several measures (including activity and popularity) that exist in literature for ranking influential users in Twitter network.

The primary contribution of this work is to highlight the expansion of influential users' posts on a Twitter graph and thus, to effectively predict, based on their influence metrics, the identification of influential users on certain subjects. Our proposed method proves that our assumption of the "influential" users playing a dominant role in information diffusion, has been verified. Furthermore, the introduction of Cluster-based Fusion of Retrieved Lists depicts the key concept of this

technique, which is that inter-similarity of documents presented in different query result lists should be rewarded. In contrast to other related works, our proposed framework utilizes the cluster fusion method, which was initially designed for document similarity ranking. This cluster fusion approach is based on the so-called "cluster hypothesis", namely, "to reward low-ranked documents with the condition that they belong in the same cluster with high ranked documents". Another contribution is the incorporation of additional features, namely time-based features (frequency, part-of-day) in the dimension of each Twitter user. Finally, another important aspect of our proposed paper is the information diffusion or else, the percentage of covering a Twitter graph achieved.

The rest of the paper is structured as follows. Section 2 focuses on 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 are depicted. Finally, Section 5 concludes our work and draws directions for future research.

2. Material and Methods

Commercial companies and associations could exploit Twitter for marketing purposes, as it provides an effective medium for propagating recommendations through users with similar interests. Moreover, viral marketers could exploit models of user interaction to spread their content or promotions quickly and widely.[27]

Recently, the identification of topical (or influential) authorities in microblogging has gained a lot of attention. In Ref. 30, 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 group 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 Refs. 18, 21 and 19, 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 Ref. 22, an efficient and innovative methodology for community detection that will also leverage users' behavior on emotional level is introduced. Also in Ref. 37, some metrics for estimating the user's Influence by presenting an analysis on the current strength of Twitter is proposed. Authors claim that usually the effect of influence is sighted when user's Followers are affected via corresponding user's posts, even though the existence of this kind of

friendship might be ignored. They utilize a two-way friendship relationship in order to produce an integrated approach of influence and this influence has been proved, through several measurements, to have bigger impact in smaller accounts (in terms of Followers), while for larger accounts, the behavior of corresponding Followers remains to be studied.

Interesting is the work presented in Ref. 35, 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 authors claim in the same work that TwitterRank outperforms other related algorithms such as In-degree, PageRank, and Topic-sensitive PageRank. Another algorithm derived from PageRank is TunkRank, which recursively computes the digital influence of Twitter account. Consequently, these influential users are so highly regarded by their respective communities that their posts become viral. Such phenomena have lately led to the need of a standardized method so as to approach and analyze their significance to online communities. Furthermore, meaningful interaction with these communities by using posts' content having specific emotional value and then analyzing its effect on users, could potentially help in gaining a deeper understanding of the users' inner workings and as a result to give an insight on human emotional interactions and conditions in general.

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,[29] the HITS algorithm proposed by Kleinberg[23] as well as their numerous variants discussed in Ref. 26. 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. In Ref. 17 authorities are identified through an induced graph enriched with account interactions.

Historically, the above as well as other approaches and techniques have been harnessed throughout microblogging areas. In Ref. 13, 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 Ref. 28, 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, in terms of other platform (e.g. in Yahoo! Answers) was addressed in Ref. 8. 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 Ref. 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.

A major contribution for understanding the importance of location and lifespan of popular discussion topics on the platform of Twitter constitutes the work presented in Ref. 5. Although this research focuses on topic popularity and distribution more than user influence, thus it provides a helpful insight on the way discussions develop through space and time on Twitter. This analysis is based on a massive dataset of over 200 million tweet involving 4000 topics of different levels of popularity. Location is treated both as real geo-location of the users contributing in the discussion topics as well as network location and proximity over the Twitter sub-graph that is produced by the active users on a topic. As most of the users in their dataset originate from the United States, the US is divided into five different time-zones. Results show that highly popular topics tend to span to more than one location/region and involve the majority of the biggest connected component of the relative sub-graph. In order to analyze the lifetime of highly popular versus not so popular topics, for each topic different sub-graphs of activity were created for every day and were compared appropriately. This lead to the helpful observation that highly popular topics keep most of the users on the biggest connected component of the sub-graph active throughout the lifetime of a topic. These insights can be projected on the popularity of users and serve as the intuition behind the proposed time-based features in the current research, which depict the engagement of influential users in ways that span access from different locations/time-zones and their involvement in terms of frequency of engagement.

In Ref. 16, the term "influencer" is introduced as a model for representing a social network, which includes two main objects, namely users and tags. In addition, influence measures that represent the ability of the influence on other people by relationships between users and the concern to user's tags as well as the speed of the user's tag propagation on the social network are identified. A recent survey[7] focuses on top-k nodes that are the important actors for a subjectively determined topic in a social network. Authors review and classify existing literature on top-k nodes identification into two major categories, namely top-k influential and top-k significant nodes. A novel methodology, called Personalized PageRank, integrating both the information obtained from network topology and the information obtained from user actions and activities in Twitter, is introduced in Ref. 3. Experimental results on a large dataset show that using user specific features like topical focus rate,

activeness, authenticity and speed of getting reaction on specific topics positively affects identifying influencers and lead to higher information diffusion.

Authors in Ref. 34 build a large-scale directed interaction graph of Twitter users and present an analysis of the geographical characteristics of the edges in this interaction graph. Three versions of PageRank that measure spatial influence on the interaction graph are proposed; Edge-Local PageRank (ELPR), which takes into account the spatial locality of edges, Source-Vertex-Locality PageRank (SVLPR), which takes into account the spatial locality of source vertices in edges, and Geographical PageRank (GPR), which incorporates the two above factors. Authors in Ref. 12 present the first known streaming algorithm for computing the H-index of a user in the cash register streaming model as they study how to calculate this metric on large streams of user publications and feedback. Publication settings with positive user feedback, such as, users publishing tweets and other users retweeting them, friends posting photos and others liking them or even authors publishing research papers and others citing these publications, were considered.

Finally, relative study with the current one is the one proposed in Ref. 4, where authors investigate whether potential similarity in the characteristics of two users can affect the evaluation that one user provides to another. Concretely, authors 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 lies 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 were evaluated.

3. Proposed Approach

In this section, the proposed data mining system in the context of social media platform is introduced. Since our motivation stems from the fact that we are interested in identifying the more influential or authoritative users per topic, a set of features need to be extracted.

The list of extracted features includes text similarity measures, social impact through retweets, the ability to spike conversations considering the content provided (through conversational tweets), as well as social graph relations. Specifically, the feature set is presented in detail along with our novel proposal to include time-based features as well as different ranking methodologies implemented.

3.1. *System architecture*

In the social media mining system we developed, the most authoritative users per topic are identified based on a variety of features that take into consideration different aspects of their Twitter dimension. 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.

Furthermore, the following Fig. 1 presents the architecture of the system which has been developed, in order to address the needs of the present work, as it illustrates its components as well as the information flow between them.

Our system architecture consists of the following phases:

- *Twitter Access*: Twitter database is accessed through Twitter API by this module, using the Twitter4j Java library for Twitter application development.

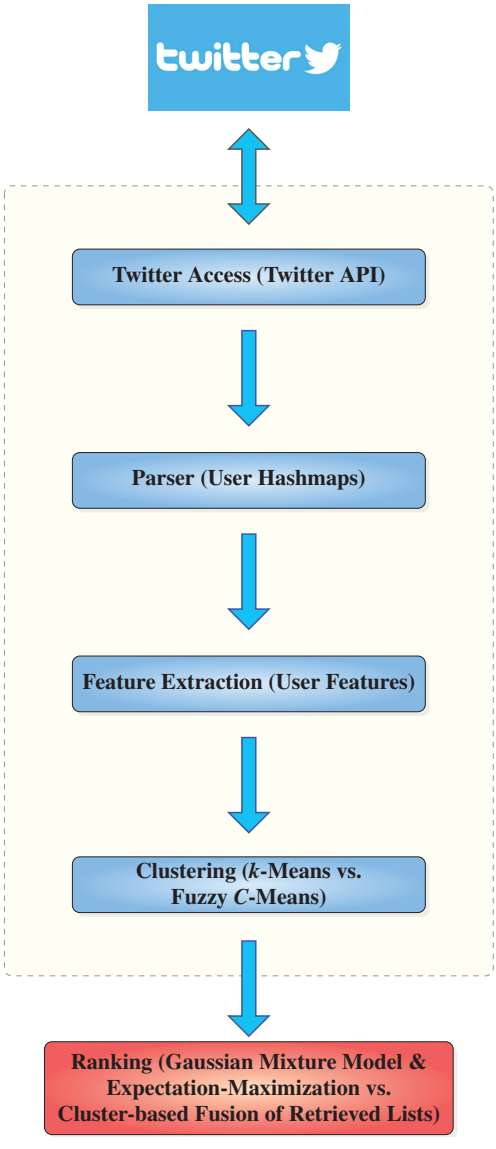

Fig. 1. Overall architecture of the proposed methodology.

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.

- *Parser*: 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*: 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 6-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.
- *Clustering*: The set of username — feature values hashmaps is given as input in a module responsible for the clustering algorithms. In this paper, data clusters are created by using two different clustering algorithms, namely k-Means as well as Fuzzy C-Means.
- *Ranking*: Different types of ranking techniques are compared at the clustered user data. Gaussian ranking used by Ref. 30 is tested against a method described in Ref. 25.

Direct access to various Twitter topics 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.

In terms of performance, the time complexity of computing the features for a user from the dataset is $O(n)$, where n is the number of tweets. In addition, the cost is $O(n^2)$, where n in this case is the number of followers or friends. With enough memory space, this system can operate 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.

3.2. *Twitter access and parser*

In this subsection, the methodology for estimating the influence of a Twitter user in a specific Twitter graph is introduced as in Ref. 21. The social media crawler creates a social media graph where the nodes represent the users and the edges represent the "Following" relationships among these users. In our paper, we utilize

Algorithm 1 Generation of Social Media Graph

Require: Query/Keyword $\#q$

Ensure: The sample Graph $Users$ and lists Followers[] and Newnodes[] are computed

identify set of tweets for given $\#q$, $T = \{t_1, t_2, \ldots, t_i\}$
$\forall$ tweet $t_i \in T$
$u_i \leftarrow$ user of tweet t_i
Followers$[i] \leftarrow$ Followers of $c\,[i]$
for all $t_i \in T$ **do**
  $Users \leftarrow Users \cup u_i$
end for
identify set of followers of a user u_k, Followers$[u_k] = \{f_1, f_2, \ldots, f_j\}$
for all $u_k \in$ Users **do**
  for all $f_j \in Followers[u_k]$ **do**
    if $f_j \in$ Users **then**
      link f_j with u_k
    else
      for all $u_l \in$ Users **and** $u_l \neq u_k$ **do**
        if $f_j \in$ Followers$[u_l]$ **then**
          Newnodes $\leftarrow$ Newnodes $\cup\, f_j$
          link f_j with u_k **and** link f_j with u_l
        end if
      end for
    end if
  end for
end for
Users = Users $\cup$ Newnodes

a topic-based sampling approach where tweets are collected via a number of different keyword search queries.

More specifically, the Twitter graph is generated in the following way. Initially, for a concrete #hashtag, the users and their corresponding followers, which have posted a tweet within a given time period, are retrieved. Subsequently, users that follow each other or have a common follower, are connected in order for the graph to be utilized. The process for generating the Social Media Graph is presented in a more analytical and detailed way in Algorithm 1.

3.3. *Feature extraction*

This subsection describes the set of features we inherited from Ref. 30 (named "Basic Features") and our contribution to the feature set, which is named "Time-based Features".

3.3.1. *Content and graph features*

A set of features per user is implemented by using user-specific measurements and in following combining them as proposed in Ref. 30. These measurements deal with the number and the content of Original Tweets (new content provided by the user on a topic), Conversational Tweets (replies to other users, signified by the "@username" string), Repeated Tweets (original content by the given user which is then reproduced by other users, signified by the "RT" string), Mentions (unique references to a user by other users) and Graph Characteristics (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) measures the percentage of author contribution in a topic, regardless of the type of tweets.
- *Signal Strength* (SS) is the ratio of original content in an author's topical signal.
- *Non-Chat Signal* ($\sim CS$) 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) counts 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 the 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 Ref. 30.

3.3.2. *Time-based features*

A central point of our proposed work constitutes the introduction of time-based features. Our motivation is to track the activity of a user throughout the lifecycle of a given topic and examine if and how this aspect contributes in user authority

and influence. Most of the research in the field disregards this aspect and focuses only in content and social relations deduced from the social graph of the users on a given social media platform. This leads to a static view of the data around a topic, showing indifference for temporal distribution, that is the way that discussion data is spread through time. The reality is that social media topic discussion is more dynamic and as a results, user engagement differs though time, especially when a topic corresponds to an emerging event.

Our experience in social media platform identifies sparks of "discussion traffic" that can be recognized when the topic is very relevant, meaning that at some time intervals, due to events of conjuncture, many users get attracted by the specific topic. Some users contribute to a topic for a short period of time while others keep providing content and discussion throughout the lifetime of a topic. Our claim is that a strongly authoritative user should provide content or be conversationally active throughout the total lifecycle of a concrete topic.

For topics with international significance and participation, user content must be discoverable in different periods of the lifecycle of a topic for a user to be considered as authoritative. Since online discussion in Twitter is stream-like, frequent participation leads user content to be read and discovered by users independently of the time they access this topic. This observation leads to our second claim that authoritative user tweets should be discoverable throughout the day so that active, in different time zones, users could interact with the authoritative user content. This is true especially for topics with a lifecycle that lasts days or even 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* constitutes a feature indicating the contribution of a specific author in a topic during the entire lifecycle of the topic. It is measured as the number of tweets per user divided by the duration of a topic. In our approach, tweet frequency is linear to the authority of the author. As argued before and in contrast to the streaming and bursting nature of the information on social networks, we claim that for a user to be more authoritative, their content generation must follow and span a large segment of the topic lifecycle. In addition, the score of users that contribute a lot in a short amount of time should be evened out and high frequency scores can rule out effects of posting burst. In the example which motivates the research in Ref. 30, that is *Gulf of Mexico Oil Spill*, Twitter accounts of environmental agencies that were considered as authoritative ones 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 following ratio is used for every author active in the topic:

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

- *Part-of-day measure* counts the average number of tweets a user posted in each of the four 6-hour parts-of-day measures (morning, noon, evening, night). Since we use four parts-of-day slots, this is an approximation feature; it 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, for a user in East Asia it is practically impossible to read original content from an author posting from the United States, unless this user covers more parts of day (taking into account that most users are not 24/7 online). Ideally, an author (such as an account registered by an institution or a news agency) should have a posting distribution that covers all day and we support that our approximation is enough to demonstrate such distribution. In each part-of-day, the average number of tweets is calculated and provided as a clustering dimension.

3.4. *Clustering and ranking*

For the clustering and ranking process, two methods were compared in order to derive possible authoritative users:

- clustering and ranking with the use of Gaussian Mixture Models (GMM) and the Expectation-Maximization (EM) algorithm,[30] and
- our proposal, that is clustering and ranking with the use of cluster-based fusion of retrieved lists.[25]

In our proposal, we have also employed, except simple k-Means algorithm, the Fuzzy C-Means algorithm,[6,10] that is as Fuzzy C-Means points out the notion of similarity, which is well suited when one has to deal with user content on a specific topic.

3.4.1. *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:

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

where x is a D-dimensional data vector of features, w_i 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 authoritative users should have frequent multi-type contribution on the topic.

3.4.2. *Cluster-based fusion of retrieved lists*

The technique of cluster-based fusion is presented and evaluated in Ref. 25. 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, the CombMNZ and the Borda methods use partial list rankings to build a final result list, which can lead to very low total ranking,[25] 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.

The motivation of the cluster-based fusion methods is better analyzed in Ref. 24. This research centralizes the concept of fusion of different result lists, especially in the context of "recommendation discovery" or "expert identification" from a given query on a set of documents, a context similar to the identification of influential users. Different result lists are produced by slightly altering the initial query on the same systems or by trying the same query on different search methodologies. The cluster-based fusion method tries to overcome the linear reward system that exists in the previous fusion methods which gives a high aggregate rank to identical documents that are already ranked highly in the individual lists. The way to implement that is to utilize the notion of the cluster hypothesis in order to identify, to group together as well as to give a higher rank to documents that have inter-document similarities and appear in different rankings among the individual lists. In other words, a support system between the different lists is created and in following, the

corresponding results have shown that the final ranking produced by this method is evaluated as more relevant one.

Three methods of list fusion are examined in a straightforward application versus an application in combination with two clustering algorithms (e.g. simple nearest-neighbor algorithm like k-Means and Fuzzy C-Means). More specifically, CombSUM method simply sums the document retrieval scores across the lists, CombMNZ enhances that by multiplying each sum by the number of lists each document appears on. Finally, the Borda method follows a different approach which is that each document is scored by the number of documents not ranked higher than it in the lists.

In our proposal, we utilize this method by comparing k-Means with the Fuzzy C-Means algorithm for clustering documents. More specifically, for the case of C-Means algorithm, the results are initially clustered into k lists, 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[24,25] best-performing versions of the algorithm.

4. Experimental Evaluation

In the next three subsections, the experimental setting for our approach is unfolded (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) for the two clustering algorithms.

The initial experiments have been recently repeated with three larger datasets, in order to validate the previous results and observe the scalability of the proposed algorithm compared to Ref. 11. Again, tweets from three different subjects were aggregated together with follower and friend relationships of the active users on each thread. This was done by using the Twitter API and taking into account its limitations.

4.1. *Twitter datasets synopsis*

For the construction of our test dataset, we had to respect the current limitations of the Twitter API, together with the need to build a dataset of topics that have differences in their temporal development. We implemented the methodology described using Twitter4j,[a] a Java based platform utilized for interacting with the Twitter API. The Twitter subgraphs, the algorithmic output and the user evaluation of

[a]http://twitter4j.org/en/index.html

the results were collected in a time interval of one month, that is (23/10/2018–23/11/2018). A topic-based sampling approach was used where tweets are collected via a keyword search query. More specifically, data for six given discussions on Twitter, namely #blacklivesmatter, #bigdata, #germanwings, #WorldSeries18, #MAGAbomber and #node.js, were downloaded.

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, reflects a discussion topic with mostly scientific and business interest and quite sparse but also quite linear activity in time. The third hashtag, #germanwings, deals 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.

Regarding the new topics, the following aspects can be considered. The next hashtag, #WorldSeries18, is a sports-related topic which runs throughout the year and has several spikes of activity responding to related sports events, the discussion before them, live discussion and the discussion that follows each event. Between the events, discussion remains in low frequency. The following hashtag, #MAGAbomber, was an emerging discussion around a mail bombing event located in the United States that attracted great public interest for a short period of time and lead to a big spike of tweets for almost a week, followed by a long tail of low activity afterwards. Finally, hashtag #node.js, is a long-running topic of discussion related with a technology that has become extremely popular during the last years and is continuously discussed by a relatively large community of software development professionals, organizations and companies. This discussion includes tips and tricks, related conference events, public updates etc.

The construction of the dataset was a two-step process by harvesting the related tweets per topic and then querying Twitter to get the followers and friends for each active user. Concretely, it was completed via a two-step repetitive process where initially a tweet was returned as answer to the hashtag query and in following a second step was performed in order to fetch the friends' as well as the followers' list of the user that initially posted the tweet. That process was followed for the total of six hashtags and resulted in the following Table 1.

Table 1. Dataset features.

Topic	Number of Tweets	Number of User Accounts	Number of Followers
#blacklivesmatter	20.146	275	350.622
#bigdata	24.569	328	398.349
#germanwings	17.945	196	286.002
#WorldSeries18	27.563	371	457.126
#MAGAbomber	18.234	178	236.544
#node.js	25.147	264	336.412

1960010-16

Table 2. Subgraph properties for six datasets.

	Value					
Property	#blacklivesmatter	#bigdata	#germanwings	#WorldSeries18	#MAGAbomber	#node.js
Vertices	350.622	398.349	286.002	457.126	236.544	336.412
Edges	545.567	632.832	437.674	789.136	499.624	670.548
Triangles	31	446	286	654	343	599
Squares	60	77	59	99	65	79
Stars	40	44	41	65	47	60
Components	50	52	46	55	44	49
Diameter	12	16	12	16	14	15
Tweets	20.146	24.569	17.945	27.563	18.234	25.147
Retweets	8.134	8.673	6.314	9.255	6.854	8.694
Avg. Tweets	156.3	67.9	87.8	99.8	106.2	55.3
Avg. Following	5.45	6.87	9.75	6.23	5.14	5.54
Avg. Followers	5.96	6.99	11.67	7.67	5.23	7.34

The overall properties of the six datasets are presented in Table 2. The first column has fundamental graph structure properties such as the number of edges and triangles as well as Twitter specific properties like the average tweet length and the average number of followers. Note that the vertices are accounts and the directed edges represent "following" relationships.

On the basis of the values of Table 2, it can be argued that all the datasets are coherent, as they contain a small number of stars and it is more closely connected as shown by the low diameter. Additionally, the tweets' length is long enough, which may imply that persons are attempting to explain their views, support friends, or rebut opposing arguments. These structural characteristics indicate an active social network.

Since our motivation stems from the fact that we are interested in creating some densely connected social media graphs and not just some random graphs, we tried to introduce the same properties in all the corresponding datasets.

4.2. *Top-k users*

For each topic and each tweet on the dataset, 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 Ref. 30 and three versions of cluster-based fusion algorithms using the ClustFuseCombMNZ, ClustFuseBorda and ClustFuseCombSUM strategies for list fusion, as presented in Ref. 25, with and without the addition of the proposed temporal features.

In the following Tables 3 and 4, the top-5 users with and without temporal features taken into account are presented. It can be observed that although most of the users appear on both tables, there are some important differences related on the introduction of the temporal features and the sensitivity of the methods based on the cluster hypothesis.

More to the point, in the four columns of Table 3, one can observe the top-5 ranked user lists of the four different algorithms for the six different topics taking as input the temporal features. On the contrary, the four columns of Table 4 present the same results but without taking into account the temporal features. 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

Table 3. 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
	#WorldSeries18		
WolfeIII	MrJamesFortun_	angkorwattourg1	MrJamesFortun
Commandpost5	angkorwattourg1	MrJamesFortun_	angkorwattourg1
LennyMarshella	Commandpost5	NUnwind	PeggyBinette
EbrTravel	NUnwind	EbrTravel	Melinda15858273
PeggyBinette	LennyMarshella	WolfeIII	NUnwind
	#MAGAbomber		
Sunstoned2	WildCathRN	Advancedape2E	GosiaZna
TerriNich299	Sunstoned2	Frankie1654	JKorinetz
WildCathRN	DesireeTrail1	georgia7077	anneottley
Bselected	anneottley	JKorinetz	TerriNich299
DoremusJ	Advancedape2E	JAngello85	Advancedape2E
	#node.js		
dev_dsgn	JavascriptFlux	CalMaths	benjaChomin
flowroute	dev_dsgn	pushkar_nk	pushkar_nk
GitLit000	walkingriver	anathijay	RayAssociatesUK
pushkar_nk	GitLit000	Bwetea	sand_9999
RayAssociatesUK	benjaChomin	flowroute	mvelosop

forms. A special occasion web application was developed linked to a database where answers were concentrated for later process. The evaluation scenario complied with the following assumptions:

- evaluating users were anonymous (age and gender data where recorded for statistical reasons) and

Table 4. 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
	#WorldSeries18		
angkorwattourg1	WolfeIII	EbrTravel	LewGarfinkel
Commandpost5	Bwetea	LennyMarshella	NUnwind
EbrTravel	LennyMarshella	LewGarfinkel	PeggyBinette
LennyMarshella	angkorwattourg1	Melinda15858273	WolfeIII
LewGarfinkel	Commandpost5	MrJamesFortun_	angkorwattourg1
	#MAGAbomber		
Advancedape2E	WildCathRN	JKorinetz	GosiaZna
anneottley	DesireeTrail1	Frankie1654	JAngello85
DesireeTrail1	Advancedape2E	keygirl24	JKorinetz
Frankie1654	JAngello85	anneottley	johnburnsnc
georgia7077	lesliejoan58	reality_UsExPat	WildCathRN
	#node.js		
anathijay	dev_dsgn	JavascriptFlux	GitLit000
benjaChomin	flowroute	GitLit000	JavascriptFlux
CalMaths	GitLit000	oss_js	mvelosop
dev_dsgn	JavascriptFlux	dev_dsgn	oss_js
flowroute	mvelosop	CalMaths	pushkar_nk

- 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 dataset and were enabled to browse through the tweets, in following to filter them by topic and finally to query them by keyword or by username. After browsing through the dataset, users were asked to

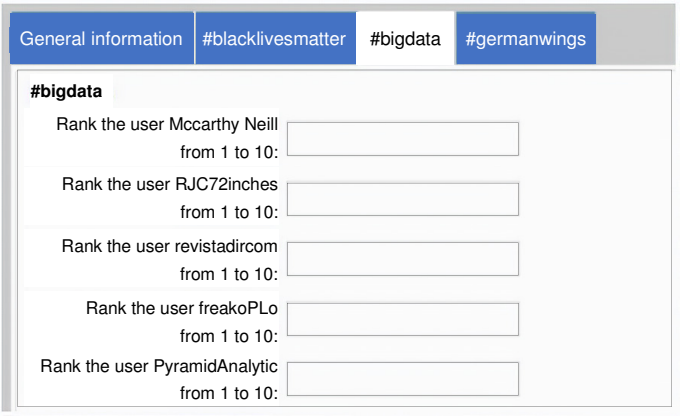

Fig. 2. Web form for user evaluation.

choose the most influential username per topic, according to their personal beliefs. A figure showing the aforementioned details is presented in Fig. 2. That username was awarded by 10 additional points. After choosing the top username, users were presented with six 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 344 social media users from Facebook and Twitter took part in the evaluation survey with average age of 27.1 years and 43% of them were women, 41% of them were men and 16% did not answer this question. 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 5 to 8 for the sets of experiments described in Subsection 4.2. Pearson-correlation metric, or Pearson's r, is presented in following Eq. (3).

$$r_{X,Y} = \frac{E[XY] - (E[X]E[Y])}{\sqrt{E[X^2] - (E[X])^2}\sqrt{E[Y^2] - (E[Y])^2}} \tag{3}$$

where X and Y are random variables, and E is the expectation (or the expected value) of a random variable. Please notice that Pearson's r is a measure of the linear correlation between two variables X and Y. It can take values in the range between -1 and $+1$, where -1 constitutes a total negative linear correlation, 0 can be considered the case where there is no linear correlation, and $+1$ depicts a total positive linear correlation.

As we can observe in both situations, the cluster-based methods score better than the GMM-based algorithm. The GMM-based algorithm seems to outrun the

Table 5. Precision with temporal features.

Method	Dataset					
	#blacklivesmatter	#bigdata	#germanwings	#WorldSeries18	#MAGAbomber	#node.js
			k-Means			
GMM	0.6	0.5	0.45	0.45	0.55	0.6
ClustFuseCombMNZ	0.5	0.5	0.5	0.6	0.5	0.6
ClustFuseBorda	0.7	0.7	0.65	0.7	0.7	0.6
ClustFuseCombSUM	0.7	0.7	0.65	0.7	0.7	0.7
			C-Means			
GMM	0.7	0.6	0.6	0.6	0.7	0.7
ClustFuseCombMNZ	0.6	0.6	0.5	0.6	0.55	0.65
ClustFuseBorda	0.85	0.8	0.75	0.85	0.8	0.7
ClustFuseCombSUM	0.8	0.8	0.75	0.85	0.75	0.7

Table 6. Precision without temporal features.

Method	Dataset					
	#blacklivesmatter	#bigdata	#germanwings	#WorldSeries18	#MAGAbomber	#node.js
			k-Means			
GMM	0.6	0.55	0.6	0.55	0.6	0.65
ClustFuseCombMNZ	0.5	0.5	0.5	0.45	0.5	0.5
ClustFuseBorda	0.7	0.65	0.6	0.7	0.6	0.65
ClustFuseCombSUM	0.65	0.7	0.65	0.7	0.7	0.65
			C-Means			
GMM	0.7	0.6	0.7	0.6	0.65	0.7
ClustFuseCombMNZ	0.5	0.6	0.5	0.55	0.6	0.6
ClustFuseBorda	0.8	0.8	0.7	0.8	0.7	0.75
ClustFuseCombSUM	0.8	0.85	0.7	0.75	0.75	0.7

Table 7. Pearson-correlation with temporal features.

Method	Dataset					
	#blacklivesmatter	#bigdata	#germanwings	#WorldSeries18	#MAGAbomber	#node.js
			k-Means			
GMM	0.4	0.43	0.46	0.46	0.45	0.45
ClustFuseCombMNZ	0.43	0.43	0.42	0.41	0.45	0.45
ClustFuseBorda	0.52	0.55	0.49	0.53	0.52	0.54
ClustFuseCombSUM	0.5	0.59	0.56	0.55	0.5	0.55
			C-Means			
GMM	0.45	0.49	0.51	0.51	0.48	0.46
ClustFuseCombMNZ	0.47	0.47	0.48	0.47	0.52	0.48
ClustFuseBorda	0.57	0.62	0.55	0.61	0.56	0.58
ClustFuseCombSUM	0.55	0.64	0.59	0.58	0.52	0.65

Table 8. Pearson-correlation without temporal features.

Method	Dataset					
	#blacklivesmatter	#bigdata	#germanwings	#WorldSeries18	#MAGAbomber	#node.js
			k-Means			
GMM	0.4	0.42	0.44	0.47	0.43	0.46
ClustFuseCombMNZ	0.4	0.4	0.44	0.44	0.45	0.43
ClustFuseBorda	0.5	0.52	0.5	0.53	0.5	0.55
ClustFuseCombSUM	0.49	0.62	0.55	0.52	0.45	0.6
			C-Means			
GMM	0.43	0.46	0.51	0.53	0.46	0.48
ClustFuseCombMNZ	0.44	0.42	0.48	0.48	0.51	0.47
ClustFuseBorda	0.58	0.57	0.55	0.58	0.54	0.61
ClustFuseCombSUM	0.52	0.66	0.59	0.55	0.48	0.67

cluster-based fusion method only when ClustFuseCombMNZ strategy is used for fusion.

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.

The results on the new datasets show that the proposed method can be scalable and affect larger data samples of Twitter activity. All four different methods seem to benefit in their Precision and Pearson-correlation metrics by the introduction of temporal features. The ClustFuseBorda method is proven to be the most effective, while the GMM method has better results than ClustFuseMNZ in both cases and matches the results of ClustFuseCombSUM when temporal features are not present.

In addition, in all Tables 5 to 8, results from Fuzzy C-Means clustering algorithm outperform traditional k-Means. This is something expected as Fuzzy C-Means points out the notion of similarity, which is well suited when one has to deal with user content on a specific topic.

4.4. *Information diffusion*

Another important output of our proposed paper is the information diffusion that considers graph patterns. In this subsection, the percentage of covering a Twitter graph is measured. When taking into account network metrics in information diffusion, a rather realistic information diffusion process can be utilized. The most important factor which affects the transmission of the tweets is the followers' probability of retweeting.[20]

Figure 3 presents the percentage of graph cover or else the rate of covered users in case the diffusion starts from the "influential" users for the #germanwings

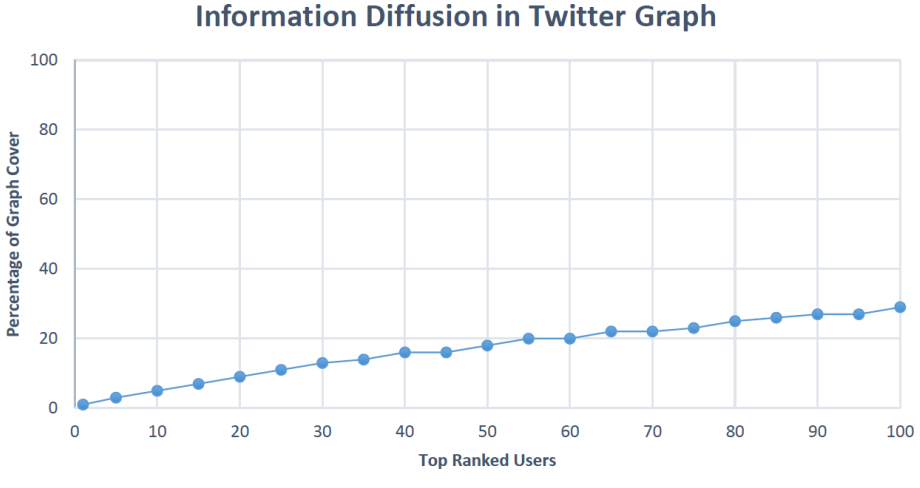

Fig. 3. Rate of covered users.

dataset. The horizontal axis represents the users in descending order, starting from the top ranked ones and the vertical axis represents the rate of covered users (or the percentage of graph cover). This percentage is derived from the "Following" relationships among users of the generated graph.

We can therefore observe, that the top 100 ranked users of the #germanwings dataset are enough in order to cover almost the 30% of the whole graph of users. Hence we can state that our assumption of the "influential" users playing a dominant role in information diffusion, is verified. To be more specific, about 30% of the whole 286 002 users regarding the corresponding dataset, "follow" one or more out of the top 100 ranked users and thus, can be informed for a specific post that they could make.

5. Conclusions and Future Work

In this paper, a novel methodology for discovering topical influential users in a microblogging environment, was presented and in following evaluated. The important advances of this research constitute 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 commonly found in social network environments. Our proposed method proves that the influential users play a dominant role in information diffusion.

As work to come, we are interested in parallelizing the methods presented in the proposed 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[32] could also expand the aspect of temporal dynamics in such environments. On the other hand, in order for the discovery of influential users to be more accurate, the properties of the microblogging network and the behavior of the users, such as understanding collaborative behavior,[14] analyzing why a tweet is likely to be retweeted[33] and decoding the social mechanism that explains why users with many followers are not necessarily the most influential,[9] can be comprehended.

References

1. F. Abel, Q. Gao, G.-J. Houben and K. Tao, Analyzing temporal dynamics in Twitter profiles for personalized recommendations in the social web, in *3rd Int. Web Science Conf.* (*WebSci*) (2011), pp. 2:1–2:8.
2. E. Agichtein, C. Castillo, D. Donato, A. Gionis and G. Mishne, Finding high-quality content in social media, in *Int. Conf. on Web Search and Data Mining* (*WSDM*) (2008), pp. 183–194.
3. Z. Z. Alp and S. G. Ögüdücü, Identifying topical influencers on Twitter based on user behavior and network topology, *Knowledge-Based Systems* **141** (2018) 211–221.

4. A. Anderson, D. P. Huttenlocher, J. M. Kleinberg and J. Leskovec, Effects of user similarity in social media, in *5th Int. Conf. on Web Search and Data Mining* (*WSDM*) (2012), pp. 703–712.
5. S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy and S. Triukose, Spatio-temporal and events based analysis of topic popularity in Twitter, in *22nd ACM Int. Conf. on Information and Knowledge Management* (*CIKM*) (2013), pp. 219–228.
6. J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* (Springer, 1981).
7. R. Bian, Y. S. Koh, G. Dobbie and A. Divoli, Identifying top-k nodes in social networks: A survey, *ACM Computing Surveys* (*CSUR*) **52**(1) (2019) 22:1–22:33.
8. M. Bouguessa, B. Dumoulin and S. Wang, Identifying authoritative actors in question-answering forums: The case of yahoo! answers, in *14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (*KDD*) (2008), pp. 866–874.
9. M. Cha, H. Haddadi, F. Benevenuto and P. K. Gummadi, Measuring user influence in Twitter: The million follower fallacy, in *4th Int. Conf. on Weblogs and Social Media* (*ICWSM*) (2010).
10. J. C. Dunn, A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters, *Journal of Cybernetics* **3** (1974) 32–57.
11. A. Georgiou, A. Kanavos and C. Makris, Finding influential users in Twitter using cluster-based fusion methods of result lists, in *14th Int. Conf. on Artificial Intelligence Applications and Innovations* (*AIAI*) (2018), pp. 14–27.
12. P. Govindan, M. Monemizadeh and S. Muthukrishnan, Streaming algorithms for measuring H-impact, in *36th ACM SIGMOD-SIGACT-SIGAI Symp. on Principles of Database Systems* (*PODS*) (2017), pp. 337–346.
13. J. Guo, S. Xu, S. Bao and Y. Yu, Tapping on the potential of Q&A community by recommending answer providers, in *17th ACM Conf. on Information and Knowledge Management* (*CIKM*) (2008), pp. 921–930.
14. C. Honeycutt and S. C. Herring, Beyond microblogging: Conversation and collaboration via Twitter, in *42nd Hawaii Int. Conf. on System Sciences* (*HICSS*) (2009), pp. 1–10.
15. B. A. Huberman, D. M. Romero and F. Wu, Social networks that matter: Twitter under the microscope, *First Monday* **14**(1) (2009).
16. T. Huynh, I. Zelinka, X. H. Pham and H. D. Nguyen, Some measures to detect the influencer on social network based on information propagation, in *9th Int. Conf. on Web Intelligence, Mining and Semantics* (*WIMS*) (2019), pp. 18:1–18:6.
17. P. Jurczyk and E. Agichtein, Discovering authorities in question answer communities by using link analysis in *16th ACM Conf. on Information and Knowledge Management* (*CIKM*) (2007), pp. 919–922.
18. E. Kafeza, A. Kanavos, C. Makris and D. K. W. Chiu, Identifying personality-based communities in social networks, in *Legal and Social Aspects in Web Modeling* (*LSAWM*) (Keynote Speech) in conjunction with the *Int. Conf. on Conceptual Modeling* (*ER*) (2013), pp. 7–13.
19. E. Kafeza, A. Kanavos, C. Makris, G. Pispirigos and P. Vikatos, T-PCCE: Twitter personality based communicative communities extraction system for big data, *IEEE Transactions on Knowledge and Data Engineering* (*TKDE*) **1**(1) (2019).
20. E. Kafeza, A. Kanavos, C. Makris and P. Vikatos, Predicting information diffusion patterns in Twitter, in *10th Int. Conf. on Artificial Intelligence Applications and Innovations* (*AIAI*) (2014), pp. 79–89.

21. E. Kafeza, A. Kanavos, C. Makris and P. Vikatos, T-PICE: Twitter personality based influential communities extraction system, in *IEEE Int. Congress on Big Data* (2014), pp. 212–219.
22. A. Kanavos, I. Perikos, I. Hatzilygeroudis and A. Tsakalidis, Emotional community detection in social networks, *Computers & Electrical Engineering* **65** (2018) 449–460.
23. J. M. Kleinberg, Authoritative sources in a hyperlinked environment, *Journal of the ACM* **46**(5) (1999) 604–632.
24. A. K. Kozorovitzky and O. Kurland, From "identical" to "similar": Fusing retrieved lists based on inter-document similarities, in *2nd Int. Conf. on the Theory of Information Retrieval* (*ICTIR*) (2009), pp. 212–223.
25. A. K. Kozorovitzky and O. Kurland, Cluster-based fusion of retrieved lists, in *34th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval* (*SIGIR*) (2011), pp. 893–902.
26. A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton University Press, 2006).
27. J. Leskovec, L. A. Adamic and B. A. Huberman, The dynamics of viral marketing, *ACM Transactions on the Web* (*TWEB*) **1**(1) (2007).
28. M. R. Morris, S. Counts, A. Roseway, A. Hoff and J. Schwarz, Tweeting is believing?: Understanding microblog credibility perceptions, in *ACM Conf. on Computer Supported Cooperative Work* (*CSCW*) (2012), pp. 441–450.
29. L. Page, S. Brin, R. Motwani and T. Winograd, The pagerank citation ranking: Bringing order to the web, Technical Report (1999), http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.
30. A. Pal and S. Counts, Identifying topical authorities in microblogs, in *4th Int. Conf. on Web Search and Data Mining* (*WSDM*) (2011), pp. 45–54.
31. F. Riquelme and P. G. Cantergiani, Measuring user influence on Twitter: A survey, *Information Processing and Management* **52**(5) (2016) 949–975.
32. G. Stilo and P. Velardi, Time makes sense: Event discovery in Twitter using temporal similarity, in *IEEE/WIC/ACM Int. Joint Conferences on Web Intelligence* (*WI*) and Intelligent Agent Technologies (*IAT*) (2014), pp. 186–193.
33. B. Suh, L. Hong, P. Pirolli and Ed H. Chi, Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network, in *2nd IEEE Int. Conf. on Social Computing* (*SOCIALCOM*)/*IEEE Int. Conf. on Privacy, Security, Risk and Trust* (*PASSAT*) (2010), pp. 177–184.
34. H. Wei, J. Sankaranarayanan and H. Samet, Measuring spatial influence of Twitter users by interactions, in *1st ACM SIGSPATIAL Workshop on Analytics for Local Events and News* (*LENS*) (2017), pp. 1–10.
35. J. Weng, E.-P. Lim, J. Jiang and Q. He, TwitterRank: Finding topic-sensitive influential Twitterers, in *3rd Int. Conf. on Web Search and Web Data Mining* (*WSDM*) (2010), pp. 261–270.
36. S. Wu, J. M. Hofman, W. A. Mason and D. J. Watts, Who says what to whom on Twitter, in *20th Int. Conf. on World Wide Web* (*WWW*) (2011), pp. 705–714.
37. V. Zamparas, A. Kanavos and C. Makris, Real time analytics for measuring user influence on Twitter, in *27th IEEE Int. Conf. on Tools with Artificial Intelligence* (*ICTAI*) (2015), pp. 591–597.