2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Semantics-enabled User Interest Detection from Twitter

Fattane Zarrinkalam *, Hossein Fani †‡, Ebrahim Bagheri‡, Mohsen Kahani * and Weichang Du †
*Ferdowsi University of Mashhad, Mashhad, Iran
† University of New Brunswick, NB, Canada
‡ Ryerson University, Toronto, Canada

***Abstract*—Social networks enable users to freely communicate with each other and share their recent news, ongoing activities or views about different topics. As a result, user interest detection from social networks has been the subject of increasing attention. Some recent works have proposed to enrich social posts by annotating them with unambiguous relevant ontological concepts extracted from external knowledge bases and model user interests as a bag of concepts. However, in the bag of concepts approach, each topic of interest is represented as an individual concept that is already predefined in the knowledge base. Therefore, it is not possible to infer fine-grained topics of interest, which are only expressible through a collection of multiple concepts or emerging topics, which are not yet defined in the knowledge base. To address these issues, we view each topic of interest as a conjunction of several concepts, which are temporally correlated on Twitter. Based on this, we extract active topics within a given time interval and determine a users inclination towards these active topics. We demonstrate the effectiveness of our approach in the context of a personalized news recommendation system. We show through extensive experimentation that our work is able to improve the state of the art.**

***Index Terms*—User Interest Detection, Social Network Analysis, Twitter, Semantic Web, Semantic Annotation.**

I. Introduction

Due to the tremendous growth of content on the Web, many service providers are now focused on customized and targeted personalization of content for their end-users. The development of techniques that can automatically detect and model users interests is an essential step towards this purpose. Social networks enable users to freely communicate with each other and share recent news, ongoing activities or views about different topics. As a result, they can be seen as a viable source of information about the users and their interests [1], [2], [3].

Current works on the identification of user interests from social networks mostly focus on Twitter because of its popularity and open access via its API. However, accurate information extraction from Twitter microposts poses unique challenges due to the special characteristics of tweets. Tweets are too short, noisy and informal and they do not provide sufficient contextual information for identifying their semantics [4], [5]. In other words, the semantics of the context of the communicated information within a tweet is often implicit.

To address these challenges, some approaches propose to enrich tweets by annotating them with unambiguous semantic concepts described in external knowledge bases such as Wikipedia/DBpedia. These knowledge bases provide explicit semantic description of concepts and their relationships. Therefore, they can provide additional contextual information about tweets and their underlying semantics [4], [5], [6]. As an example, for a tweet such as *"The opportunity to go top of the Premier League will give Arsenal an extra incentive to beat Spurs, according to Wenger http://bit.ly/chgPjO"*, one can identify four phrases for which a related semantic concept can be identified in Wikipedia: (1) Premier League, which refers to the concept *Premier League*, an English professional league for men's association football clubs; (2) Arsenal, which can be related to the concept *Arsenal F.C.*, (3) Spurs, which can be annotated with the concept *Tottenham Hotspur F.C.*; (4) Wenger, which refers to the concept *Arsene Wenger*, a French football manager and former player.

Recent works in the domain of user interest detection has already looked into using such concepts. These works consider each of the semantic concepts, separately as a topic of interest (like the concept *Arsenal F.C.*) [2], [3], [7], or infer broader interests by traversing the concept relationship hierarchies, e.g. *Football in England* would be the broader concept that covers the semantic concepts *Premier League*, *Arsenal F.C.* and *Tottenham Hotspur F.C.* [8], [9]. However, such approaches undermine the fact that a user might not be much interested in *Football in England* or *Arsenal F.C.* as a broad topic, but be rather interested specifically in the rivalry between Spurs and Arsenal, which the current models cannot support. The support for this would require an interest to be represented through a combination of multiple semantic concepts.

In addition, existing work often confine users' interests to a set of predefined semantic concepts (e.g. Wikipedia concepts), and therefore, they cannot discover emerging topics of interest that are not explicitly included in the knowledge base [2], [10], [11]. This is important since it is very usual that some concepts collectively form a new topic in the Twitter sphere in response to an event in the real world. For example, in November 2010, Jack Wilshere, England and Arsenal footballer, received a caution for common assault over a street brawl, which received much attention on Twitter. Looking at Wikipedia, there is no entry dedicated to this event. As a result, by considering only the predefined Wikipedia concepts as topics of interest, it is not possible to unambiguously and comprehensively describe this topic.

This paper seeks to address the above shortcomings by proposing a framework for the identification of user interests

978-1-4673-9618-9/15 $31.00 © 2015 IEEE
DOI 10.1109/WI-IAT.2015.182

that emerge in time. More specifically, the key contributions of our work are as follows:

1) We define a topic of interest as a conjunction of several semantic concepts which are temporally correlated on Twitter. For example the conjunction of the *Premier League*, *Arsenal F.C.*, *Tottenham Hotspur F.C.* and *Arsene Wenger* semantic concepts can form a topic of interest to represent rivalry between Spurs and Arsenal. This has the added benefit that each detected topic of interest does not necessarily need to be represented by using a single semantic concept from an external knowledge base.
2) We propose a measure to compute *ephemeral semantic correlation* between two concepts using Twitter data in a specified time interval. Based on this measure, we construct a concept graph and utilize state-of-the-art community detection methods to detect active topics of interest in a given time interval. This is in response to the fact that the relationship between two topics on a social network changes over time and therefore, it is not possible to compute the relationship between two topics by only considering the encyclopedic similarity of the concepts that form each topic.
3) We propose a technique to determine a given users' position with regards to the active topics on Twitter hence modeling their interest to the current emerging topics within the social network space. We apply our work in the context of personalized news recommendation in order to compare to the state of the art.

The rest of the paper is organized as follows: Section II reviews the related work. The problem definition and the proposed approach are introduced in Section III and Section IV, respectively. Section V is dedicated to the Experiments and evaluation of the results. Finally, Section VI concludes the paper.

II. Background Litreture

We review the work in user interest detection from social networks in three broad categories, namely *Bag of Words*, *Mixture Model* and *Bag of Concepts* approaches.

A. Bag of Words Approach

In the *Bag of Words* approach, each user interest is represented as a term extracted from the user contents. For example, Chen et al. [12] have focused on URL recommendation in Twitter by modeling the interests of each user as a bag-of-words profile considering the words that are included in her tweets and the tweets of her followers. Shin et. al. [13] have proposed a graph-based approach for detecting topics of long-term interest to a user from Tweets. They have considered each topic to be a single term and have distinguished between persistently topical terms (PT) and other keywords by introducing two characteristics for PT terms namely *regularity* and *topicality*.

Since users in social networks can freely publish posts without any restriction, terms that they use in their posts are unstructured and unlimited. Using Bag of Words approach that focus on terms suffers from the known problems in natural language processing like Polysemy and Synonymy [14]. Furthermore, Bag of Words representation forgoes the underlying semantics of the phrases in favor of highlighting the role of syntactical repetition of textual content.

B. Mixture Model Approach

In the *Mixture Model* approach, user interests are modeled as a mixture of various topics, where a topic is a set of terms extracted from the user contents. Current works related to this approach can be categorized into two major groups based on the type of algorithm they use: *topic modeling* or *feature-based* methods.

Topic modeling methods (e.g., LDA), designed for regular documents not microposts, assume that a document is a mixture of topics and implicitly use co-occurrence patterns of terms to extract sets of correlated terms as topics of a text corpus [15]. As a result, they may not perform so well on short, noisy and informal texts like tweets and might suffer from the sparsity problem [7], [16], [17], [18]. To address this problem, some works aggregate multiple short texts to create a single document and discover the topics by running LDA over this document [18], [19], [20]. For example, Weng et al. [19] have created a single document from the collection of a user's tweets. However, in social networks a small number of users account for a large amount of the content. This makes the aggregation process less effective. Some recent works are dealing with the sparsity problem by applying some restriction to simplify the conventional topic model, for instance by assigning only one topic to each short text, or proposing a novel topic model for short texts [17], [21], [22]. Applying such restrictions and the fact that the number of topics in LDA is assumed to be fixed, can be considered strong assumptions for social network content because of the dynamic nature of social networks. Further, most of these works only focus on users textual content without taking into account much valuable information like network structure of users, content interconnections and other rich attributes such as timestamps and hashtags.

Feature-based methods focus on some features of user content (e.g., tags and named entities) and apply clustering algorithms to extract sets of related terms that can form topics. For example, Sayyadi et al. [23] have built a graph of named entities based on their co-occurrence in the documents and have used a community detection method for event detection. Cataldi et al. [24] have constructed a co-occurrence graph of emerging terms selected based on both the frequency of their occurrence and the importance of the users. The authors have applied a graph-based algorithm in order to extract emerging topics. Most of these works compute similarity between pairs of terms based on their co-occurrence patterns. Petkos et al. [25] have argued that the algorithms that are only based on pair-wise co-occurrence patterns cannot distinguish between topics which are specific to a given corpus. Therefore, they have proposed a soft frequent pattern mining approach to detect finer grained topics.

Our proposed approach is closely related to feature-based methods, because it detects topics of interest using graph-based clustering techniques. However, in feature-based methods, each extracted topic is a set of terms. Therefore, it is not possible to automatically extract the underlying semantics of each topic. In contrast, we annotate tweets with concepts defined in external knowledge bases and attempt to cluster extracted concepts to detect topics of interest. In addition, feature-based methods use mostly co-occurrence patterns to compute inter-term similarity, which leads to the sparsity problem due to the short length of tweets. We compute inter-concept similarity by customizing the concepts co-occurrences within a single tweet to an increased, yet semantic preserving context.

C. Bag of Concepts Approach

There is another line of work that represents user interests as a *Bag of Concepts*. These works connect meaningful sequences of terms mentioned in textual contents to unambiguous concepts from a large knowledge base, such as DBpedia/Wikipedia, Freebase and Yago. Since these knowledge bases represent the concepts and their relationships, they can provide the means for inferring the underlying semantics of content [4], [5], [6].

For example, Abel et al. [1], [26] have proposed to enrich Twitter messages by linking them to related news articles and then extracting the named entities mentioned in the enriched messages using web services provided by OpenCalais. The enrichments are considered to form the user interests. Michelson and Macskassy [7] have proposed the extraction of Twopics by first extracting a set of Wikipedia entities from a users tweets and then identifying high-level user interests by traversing and analyzing the Wikipedia categories of the extracted entities. Kapanipathi et al. [27] have modeled users' interests by annotating their tweets with DBPedia concepts, and use these annotations to filter tweets based on the users' interests. Kapanipathi et al. [8] have also used Wikipedia category hierarchy to extract broader interests of a user using entities mentioned in the user's tweets. Orlandi et al. [3] have used Zemanta as a named entity extractor to connect users contents to DBpedia resources in order to extract the DBpedia categories associated with each tweet. They have shown that user profiles based on DBpedia resources are more accurate than the profiles based on DBpedia categories.

Existing literature in this line of work struggle with two limitations: (1) they represent each topic of interest through an individual semantic concept. Therefore, it is not possible to infer more specific topics which are only expressible by combining multiple related concepts; (2) topics of interest are confined to a set of predefined concepts, and it is not possible to identify emerging topical interests which are not yet expressed in the knowledge base. Our proposed approach is related to the Bag of Concepts approach but we address these two limitations by viewing each topic of interest as a conjunction of several semantic concepts which are temporally correlated on Twitter. Hence, we are able to extract active topics of interests in a given time interval and calculate interest of a user to each extracted topics.

III. Problem Definition

The overarching objective of our work is to identify a users interests, within a specific time interval T, towards the topics on the Twitter sphere. In this section, we concretely formulate this problem. To this end, we first provide some foundational definitions.

Definition 3.1 (Tweet) A tweet m is defined as a 5-tuple, $m = (id, text, owner, time, RT_{flag})$, where $m.id$ is a unique numerical identifier associated with the tweet, $m.text$ is its textual content, $m.owner$ and $m.time$ denote the user who posted and the creation time of the tweet m, respectively. Finally, $m.RT_{flag}$ determines whether the tweet is a retweet of another ($RT_{flag} = 1$), or an original tweet ($RT_{flag} = 0$).

Based on Definition 3.1, we can now define the set of all tweets in a time period and the associated set of posters. In a specified time interval $T = [t_k, t_{k+1}]$, the set of tweets posted during this time interval T is denoted by M^T, i.e., $M^T = \{m | t_k \leq m.time \leq t_{k+1}\}$. Further, U^T is the set of users who have created tweets of M^T, i.e., $U^T = \{m.owner | t_k \leq m.time \leq t_{k+1}\}$ and $M^T_u \subset M^T$ is the subset of tweets posted by user $u \in U^T$.

We annotate each tweet $m \in M^T$ with semantic concepts defined in Wikipedia using an existing semantic annotation system. In a specified time interval T, C^T is the set of concepts extracted from M^T. We have used TAGME [4] in our experiments. For example, for a given tweet: *"Arsenal won't win with Wenger's policy. Spurs continue to exceed expectations"*, TAGME identifies three entities and links each of them to a concept represented in Wikipedia: Arsenal is linked to the semantic concept represented in http://en.wikipedia.org/wiki/Arsenal_F.C.; Wenger is linked to http://en.wikipedia.org/wiki/Arsene_Wenger and Spurs is linked to http://en.wikipedia.org/wiki/Tottenham_Hotspur_F.C..

Definition 3.2 (Topic) Given C^T, a topic z is defined as a set of weighted semantic concepts $z = \{(c, w(c, z)) | c \in C^T\}$, where $w(c, z)$ is a function that denotes the importance of concept c in topic z.

This definition is based on the idea that we view each topic as a semantically cohesive group of concepts. Let T be a specified time interval, then $\mathbb{Z}^T = \{z_1, z_2, ..., z_K\}$ denotes a set of active topics present in the social network in time interval T.

Definition 3.3 (Interest Profile) Let T be a specified time interval and let $u \in U^T$ and $\mathbb{Z}^T = \{z_1, z_2, ..., z_K\}$, an Interest Profile of user u in time interval T, called P^T_u, is represented by a vector of weights $(i_{u,1}, i_{u,2}, ..., i_{u,K})$. Each component $i_{u,n}$ of P^T_u denotes the degree of u's interest in the topic $z_n \in \mathbb{Z}^T$ in time interval T.

In our work, an interest profile for a user is a collection of weights showing the interest of a given user with regards to the available topics in the social network in T.

Definition 3.4 (User Interest Detection) Let T be a specified time interval, Given M^T and $u \in U^T$, the goal of the User Interest Detection problem is to infer P_u^T.

We divide this problem into two subproblems: *Semantic Topic Extraction* and *User Interest Identification*, in which the output of the first subproblem becomes the input of the second one. Specifically, in the former subproblem, given M^T as input, we aim at identifying $\mathbb{Z}^T$, i.e. the topics formed in the social network in time interval T, and in the latter, given $\mathbb{Z}^T$, $u \in U^T$ and M_u^T, we are seeking to model P_u^T.

IV. Proposed Approach

In this section, we describe our proposed approach for addressing the two subproblems: *Semantic Topic Extraction* (identifying $\mathbb{Z}^T$), and *User Interest Identification* (determining P_u^T).

A. Semantic Topic Extraction

To identify $\mathbb{Z}^T$, we annotate each tweet $m \in M^T$ with concepts defined in Wikipedia using an existing semantic annotation system. We then model the extracted concepts, denoted by C^T, as a concept graph CG^T based on the following definition:

Definition 4.1 (Concept Graph) Let T be a specified time interval, given M^T, a concept graph at T, denoted $CG^T = (C^T, E_{CG})$, is a weighted undirected graph representing underlying semantics of the tweets in M^T. C^T is a collection of Wikipedia concepts extracted from M^T and E_{CG} denotes a set of edges representing the relationships between these concepts. We define the ephemeral semantic correlation as a weight function $\Phi^T(c_i, c_j)$ that assigns a weight to each edge $e \in E_{CG}$ between two concepts $c_i, c_j \in C^T$.

Here, *ephemeral semantic correlation* between two concepts determines the relationship between two concepts in a given time interval. Since the relationship, e.g. co-occurrence, between two concepts can change over time, this value cannot be computed using a Wikipedia-based measures (e.g. WLM [28]), which for instance compute the relatedness of two concepts by link structure analysis techniques over the Wikipedia pages associated with those concepts. For instance, computing the relatedness of two concepts *Arsenal F.C.* and *Tottenham Hotspur F.C.* based on Wikipedia link structure analysis results in the same value both in November 2010 and April 2010. But these concepts have appeared much more frequently on Twitter during the November 2010 time period because of the match between Spurs and Arsenal. Therefore, while the information content value that any two concepts share, lies within their semantics, but the dynamics of topics on social networks can temporarily impact the relationship between the two concepts.

To address temporal issues, we compute the ephemeral semantic correlation between two concepts based on the co-occurrence of those concepts in the tweets published in time interval T. Given the short length of tweets, we consider the context of concepts. The relatedness of two concepts c_1 and c_2, as defined by Definition 4.2, is computed based on the relatedness between the tweets which are annotated with c_1, and the tweets annotated with c_2.

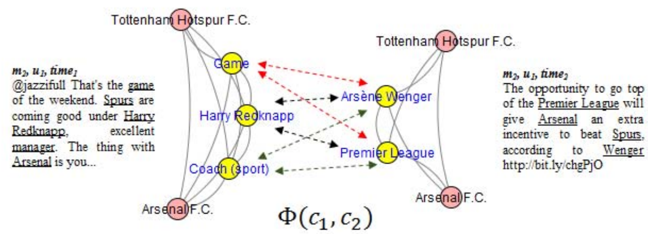

Fig. 1. Ephemeral semantic correlation: Two tweets m_1 and m_2 are from the same user u_1 posted on Nov 19, 2010.

Definition 4.2 (Ephemeral Semantic Correlation) Let T be a specified time interval and let $M_{c1}^T \subset M^T$ denotes the set of tweets which are annotated with c_1 and $M_{c2}^T \subset M^T$ denotes the set of tweets annotated with c_2, *ephemeral semantic correlation* between concepts $c_1, c_2 \in C^T$, denoted $\Phi^T(c_1, c_2)$, is defined as:

$$\Phi^T(c_1, c_2) = \frac{\sum_{m_1 \in M_{c_1}^T} \sum_{m_2 \in M_{c_2}^T} \varphi(m_1, m_2)}{|M_{c1}^T||M_{c2}^T|} \tag{1}$$

where the relatedness between two tweets m_1 and m_2, $\varphi(m_1, m_2)$, is:

$$\begin{cases} 1 - \frac{|m_1.time - m_2.time|}{|T|} & \text{if } m_1.owner = m_2.owner \\ & \text{and } m_1.RT_{flag} = m_2.RT_{flag} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The underlying idea is that if a user u has posted two subsequent tweets with some delay, e.g. less than one hour, some reasonable probability exists that the user is posting about a related subject. Therefore, the less the difference between the posting time of the two tweets m_1 and m_2 is for a given user, the greater the probability would be for those tweets to be related. For example, as illustrated in Fig. 1, two semantic concepts *Game* and *Arsene Wenger* do not co-occur in one tweet. However, since user u_1 has mentioned them in two of his subsequent tweets, they are likely to be related to each other.

Given the concept graph CG^T, we apply graph clustering and community detection methods to find semantically cohesive subgraphs of concepts as topic graphs based on the following definition:

Definition 4.3 (Topic Graph) Let $CG^T = (C^T, E_{CG})$ be a concept graph in the time interval T, a topic graph $TG = (V_{TG}, E_{TG})$ is an induced subgraph of CG^T, i.e. $V_{TG} \subset C^T$ and E_{TG} consists of those edges of CG^T with both end vertices in V_{TG}, such that its internal cohesion,

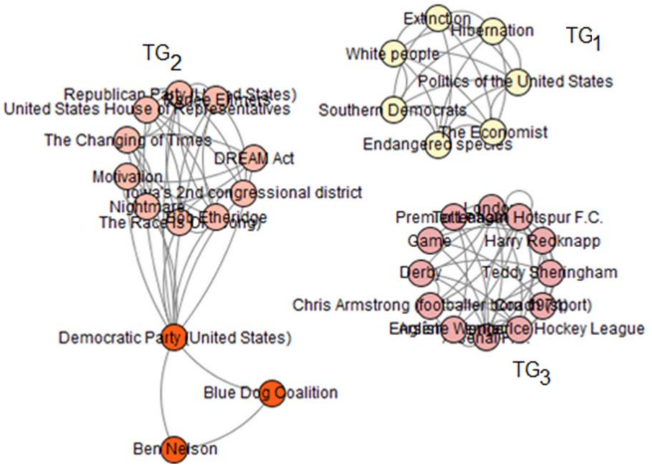

Fig. 2. Part of concept graph on tweets of Nov 19, 2010 and its topic graphs detected by VOS community detection method (Quality: 0.93, Modularity: 0.86).

denoted as $\partial_{int}(TG)$, is more than a specific threshold.

$$\partial_{int}(TG) = \frac{\sum_{\{c_1,c_2\}\in E_{TG}} \Phi^T(c_1, c_2)}{|E_{TG}|} \tag{3}$$

In Definition 4.3, internal cohesion is defined as the average *ephemeral semantic correlation* between all the pairs of semantic concepts associated with a topic graph. We will empirically show that a suitable threshold for $\partial_{int}(TG)$ can be found. Fig. 2 depicts the visualization of three topic graphs detected in Nov 19, 2010. For example, the topic graph TG_3 is formed because of the rivalry between Spurs and Arsenal. Also TG_1 refers to the discussions about the lack of a strong Democratic presence in southern states in the US after a publication in The Economist [1].

To represent each topic $z \in \mathbb{Z}^T$ based on Definition 3.2, We transform each extracted topic graph $TG = (V_{TG}, E_{TG})$ to a set of weighted concepts $z = \{(c, w(c, z))|c \in V_{TG}\}$, where $w(c, z)$ is the Degree Centrality of c in TG, computed by summing the weights attached to the edges connected to concept vertex c in the graph TG. We normalize $w(c, z)$, such that $\sum w(c, z) = 1$. Higher values for $w(c, z)$ indicate that the corresponding concept c is more closely related to topic z. Finally, all topics z from CG^T collectively form $\mathbb{Z}^T$.

B. User Interest Identification

After detecting $\mathbb{Z}^T = \{z_1, z_2, ..., z_K\}$ from the concept graph CG^T, and in order to identify $P_u^T = (i_{u,1}, i_{u,2}, ..., i_{u,K})$, we need to measure $i_{u,n}$, i.e. the interest of user $u \in U^T$ against each extracted topic z_n based on M_u^T, i.e., the tweets that the user has posted in T. Our intuition for calculating $i_{u,n}$ is that the more frequently the concepts of a topic are mentioned in the tweets of a user, the more interested

[1] http://www.economist.com/node/17467202

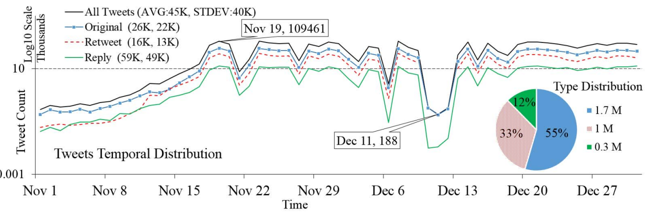

Fig. 3. Temporal distribution of different types of tweet in our dataset.

the user would be in that topic. The required $i_{u,n}$ is computed as follows:

$$i_{u,n} = \sum_{c \in z_n} w(c, z_n) * tf(c, M_u^T) \tag{4}$$

where $tf(c, M_u^T)$ denotes the frequency of the concept c in the tweets that user u has posted in T. Finally, $P_u^T = (i_{u,1}, i_{u,2}, ..., i_{u,K})$ represents the Interest Profile of user $u \in U^T$. We normalize each P_u^T, such that $\sum i_{u,n} = 1$.

V. Experiments

In this section, we describe our experiments in terms of the dataset, setup and performance compared to the state of the art. The experiments are conducted on an Intel(R) Xeon(R) 3.50GHz, 30 GB RAM.

A. Dataset

We use Abels [1] Twitter dataset including 3M tweets posted by approximately 135K users, starting from Nov 1st and lasting for two months until Dec 31st 2010. Fig. 3 depicts the overall and temporal distributions of different types of tweets. The dataset encompasses around 77K news articles crawled from the URLs mentioned in the tweets.

Social networks suffer from one common characteristic, i.e. participation inequality, where a minority of users usually contribute the most while the others just free-ride. Likewise, our dataset suffers from the same phenomenon. Fig. 4 clearly depicts that only 15% of the users contribute more than 16 tweets within a two month period and the other users have less than 16 tweets. There are 1% of users (1.5K) who actively participate by posting tweets.

We annotate tweets using the TAGME RESTful API [4] with the recommended scoring threshold of 0.1. The annotations consist of around 220K concepts conforming to a power-law distribution. Fig. 5 illustrates the 100 most frequent concepts after stop concepts such as *Retweet* and *Hypertext Transfer Protocol* are removed. The horizontal study of the concepts reveals that concepts have bursty ephemeral lifespans.

B. Evaluation Methodology

We evaluate our work by comparing it against the most relevant work in the literature by adopting their dataset and evaluation methodology. The work by Abel et al. [1] serves as a great benchmark as their dataset, and evaluation platform

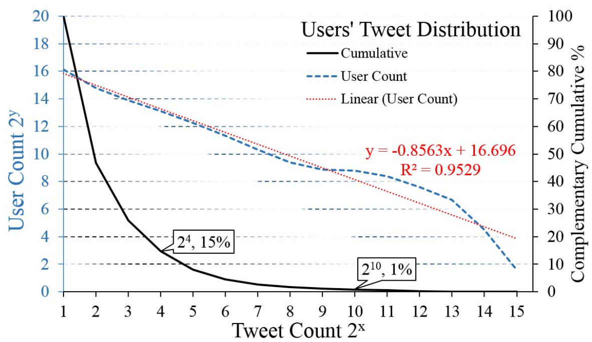

Fig. 4. The number of tweets per user and its complementary cumulative distribution.

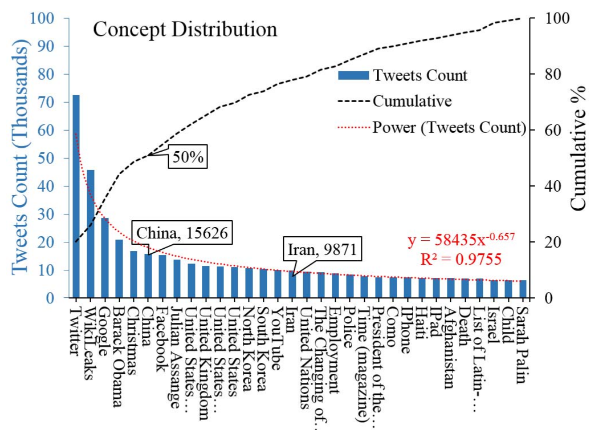

Fig. 5. Top most frequent concepts after the removal of stop concepts.

is openly available for comparison and has been widely used in the literature.

Adopted from [1], we deploy a news recommender application for evaluation purposes. First, a ground truth is built by collecting, for each user, the news articles from BBC or CNN to which the user has explicitly linked in his tweets (or retweets) by mentioning the corresponding URL in a given time interval. Then, a news recommendation algorithm is used that is able to recommend news articles based on the users interests identified by our method. By comparing the recommended news with the ones in the ground truth, it is possible to evaluate the quality of the recommendations, and therefore determine how successfully the interests of a user have been identified. Since our main goal is not to propose a news recommender system, a simple recommender algorithm is used as follows:

We represent each news article A as a weighted vector $A = (i_{A,1}, i_{A,2}, ..., i_{A,K})$ over the extracted topics $\mathbb{Z}^T = \{z_1, z_2, ..., z_K\}$. The value for $i_{A,n}$ is calculated as follows:

$$i_{A,n} = \sum_{c \in z_n} w(c, z_n) * tf(c, A) \quad (5)$$

where $tf(c, A)$ denotes the frequency of the concept c in news article A. Further, we normalize $i_{A,n}$, such that $\sum i_{A,n} = 1$.

Given A and P_u^T, it is possible to compute the interest of user u to article news A in time interval T, by calculating the cosine similarity of those vectors. Abel et al. [1] have also used a similar approach for evaluating their user interest profiles that we will compare against in this paper.

C. Evaluation Metrics

We evaluate the quality of the recommendations using standard information retrieval metrics, namely *Mean Reciprocal Rank (MRR)*, which indicates the inverse of the first position that a correct item occurs within the ranked recommendations and *Success at rank K (S@K)*, which shows the probability that at least one correct item occurs within the top-k ranked recommendations. In the experiment, the length of the time interval has been set to be 1 day. For each day of the dataset, we first compute each of these metrics for each user and then report the average of the results across all users and days.

D. Parameter Setting

As mentioned in Definition 4.3, we need a community detection method to extract topic graphs. Further, we need to empirically determine the value of the threshold for $\partial_{int}(TG)$. To do so, we have conducted an experiment in which three different clustering algorithms, namely *Strong Periodic*, *Louvain*, and *VOS*, have been separately used in our proposed approach.

Strong Periodic clustering method works on unweighted graphs and dissects the graph into strong periodic components. Instead of finding components, Louvain [29] is an efficient heuristic method that finds communities by considering both modularity and extraction time from a weighted graph. We use the work reported in [30], which is a multilevel version of Louvain with resolution parameter $r = 1$. The third method that we use, Visualization Of Similarity (VOS), provides a low-dimensional visualization of objects. It presumes distance as a similarity score and places similar objects within similar clusters.

In Fig. 6, the results of these three clustering algorithms with varying values for internal cohesion are presented. As this figure shows, exploiting VOS community detection method for extracting topic graphs improves MRR and S@10 significantly in comparison with the other two methods. Further, when the lower bound of internal cohesion is decreased from 5 to 3, the value of S@10 is increased in all methods. When the lower bound of internal cohesion decreases to values less than 3, the value of S@10 is either nearly fixed or the increase is negligible. The same discussion can be made in terms of the MRR metric which is used, in addition to S@10, in our evaluations.

On the other hand, as the lower bound of internal cohesion decreases, the number of topics and consequently the execution time is increased. This increase is noticeable when the values smaller than 3 are used as the lower bound of internal cohesion. Since smaller values for the threshold of $\partial_{int}(TG)$ increase the execution time noticeably, without improving the quality of recommendations with regards to S@10 and MRR, we select the value of 3 for the threshold.

Considering the points mentioned above, we have used VOS community detection method with the value of 3 for the threshold of its internal cohesion in our proposed approach in the rest of the experiments.

E. Comparison with baseline methods

Since our method is designed to provide improvement over Bag of Concepts approach, we evaluate our work by comparing it against two baseline methods, i.e. topic-based method and entity-based method, introduced by Abel et. al. [1]. In these methods, for a given user u, her interests are represented by a weighted vector of concepts where the weight of a concept c is equal to the number of u's tweets that refer to the concept c. In the topic-based baseline method, OpenCalais taxonomy which contains 18 topics, is used as the underlying annotations. OpenCalais is also used in the entity-based baseline method to identify entities mentioned in the tweets. While topic and entity-based methods leave 60% and 70% of tweets with no annotation, TAGME distinctly outperforms and covers 85% of the tweets with at least one high quality concept.

Fig. 7 shows that there is a positive relationship between tweet annotation coverage and user profile quantity. In other words, the more an approach extracts concepts from tweets, the less it misses the associated user profiles. Our approach which relies the annotations from the TAGME semantic annotator, outperforms the baselines' user profile coverage, increasing them from 64% and 77% to 94% as shown in Fig. 7 (right).

The results of comparing our proposed approach with the two baselines are presented in Fig. 8 in terms of both MRR and S@10 metrics. For our method, the value of S@10 is 0.504 which means that it is 50% probable that there is at least one expected result in the top-10 results of the recommendation method. This probability is about 31% and 14% for the two baseline methods. This shows that our method significantly improves the quality of recommendations in terms of S@10. The value of MRR for our method is 0.239, while the entity-based and the topic-based baseline methods have MRR values of 0.177 and 0.062, respectively. Considering the fact that the higher MRR is, the better the results of the recommendation method would be, it can be concluded that our method outperforms both entity-based and topic-based baseline with regards to MRR, pointing to a more accurate identification of user interests.

VI. Conclusion and Future Work

In this paper, we have proposed to model an emerging topic of interest on Twitter as a conjunction of several semantic concepts which are temporally correlated. To extract such topics in a specified time interval, we construct a concept graph whose nodes are a collection of Wikipedia concepts extracted from the tweets in that time interval and the relationships between these concepts are computed by measuring their ephemeral semantic correlation. Given the concept graph, we utilize state-of-the-art community detection methods to detect active topics of interest in a given time interval and determine a given users inclination towards these active topics on Twitter. We investigated the performance of the proposed approach in the context of personalized news recommendation. The

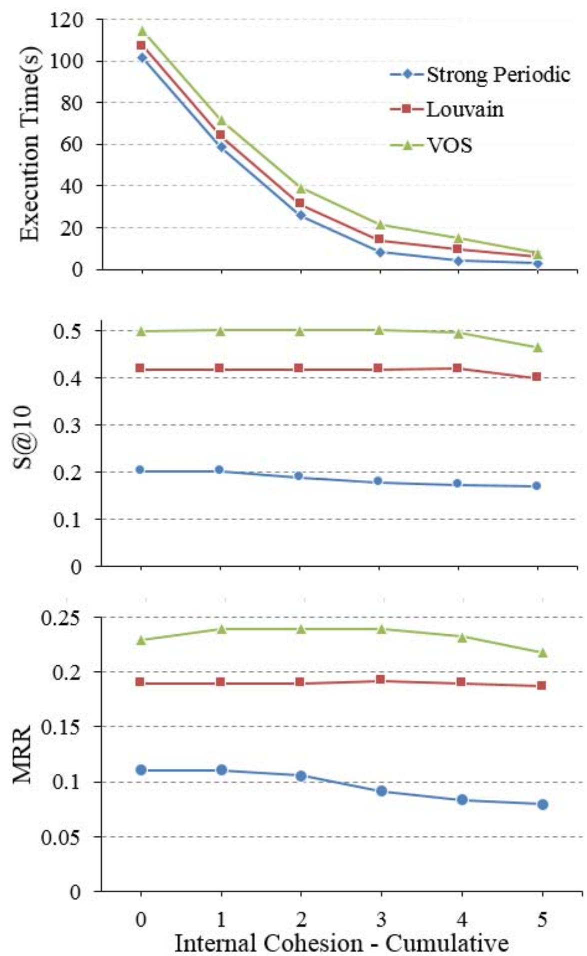

Fig. 6. Comparison between different community detection methods by varying the value of their threshold for internal cohesion in context of news recommendation.

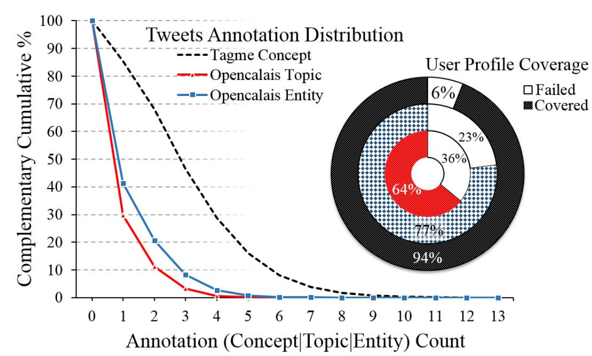

Fig. 7. Comparison of coverage between different kinds of annotations and the influence on user profile inference rate.

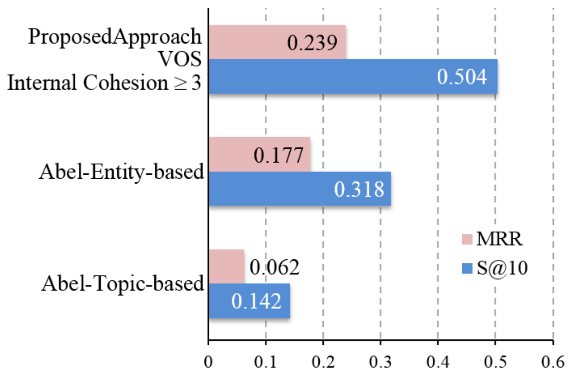

Fig. 8. Comparison between different community detection methods by varying the value of their threshold for internal cohesion in context of news recommendation.

experimental results indicate that the proposed model achieves better performance in comparison with the state of the art.

In our future work, we intend to consider other attributes of tweets (e.g., embedded URL, #hashtag and @mention) to compute the relatedness between two tweets for customizing the concepts co-occurrences within a single tweet to an increased, yet semantic preserving context. Further, the results in Section V-D indicate that choosing an appropriate community detection method has significant effect on the performance of the proposed approach. Therefore we plan to apply other types of community detection methods to better extract topic graphs.

References

[1] F. Abel, Q. Gao, G. Houben, and K. Tao, "Analyzing user modeling on twitter for personalized news recommendations," in *19th International Conference on User Modeling, Adaption and Personalization (UMAP'11)*, 2011, pp. 1–12.

[2] C. Budak, A. Kannan, R. Agrawal, , and J. Pedersen, "Inferring user interests from microblogs," in *Technical Report, MSR-TR-2014-68*, 2014.

[3] F. Orlandi, J. Breslin, and A. Passant, "Aggregated, interoperable and multi-domain user profiles for the social web," in *8th International Conference on Semantic Systems (I-SEMANTICS '12)*, 2012, pp. 41–48.

[4] P. Ferragina and U. Scaiella, "Fast and accurate annotation of short texts with wikipedia pages," *IEEE Software*, vol. 29, no. 1, pp. 70–75, 2012.

[5] P. Mendes, M. Jakob, A. Garcia-Silva, and C. Bizer, "Dbpedia spotlight: Shedding light on the web of documents," in *I-Semantics 2011*, 2011, pp. 1–8.

[6] A. Varga, E. Cano, M. Rowe, F. Ciravegna, and Y. He, "Linked knowledge sources for topic classification of microposts: a semantic graph-based approach," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 26, pp. 36–57, 2014.

[7] M. Michelson and S. Macskassy, "Discovering users topics of interest on twitter: A first look," in *4th Workshop on Analytics for Noisy Unstructured Text Data (AND'10)*, 2010, pp. 73–80.

[8] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, "User interests identification on twitter using a hierarchical knowledge base," in *11th Extended Semantic Web Conference (ESWC '14)*, 2014, pp. 99–113.

[9] C. Lu, W. Lam, and Y. Zhang, "Twitter user modeling and tweets recommendation based on wikipedia concept graph," in *AAAI 2012 Workshop on Intelligent Techniques For Web Personalization and Recommender Systems*, 2012.

[10] N. Spasojevic, J. Yan, A. Rao, and P. Bhattacharyya, "Lasta: Large scale topic assignment on multiple social networks," in *20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*, 2014, pp. 1809–1818.

[11] P. Mendes, A. Passant, P. Kapanipathi, and A. Sheth, "Linked open social signals," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010, pp. 224–231.

[12] J. Chen, R. Narin, L. Nelson, M. Bernstein, and E. Chi, "Short and tweet: experiments on recommending content from information streams," in *28th international conference on Human factors in Computing Systems (CHI '10)*, 2010, pp. 1185–1194.

[13] C. R. Y. Shin and J. Park, "Automatic extraction of persistent topics from social text streams," *World Wide Web*, vol. 17, no. 6, pp. 1395–1420, 2013.

[14] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," *Artificial Intelligence Research*, vol. 34, no. 1, pp. 443–498, 2009.

[15] D. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[16] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 841–842.

[17] C. Xueqi, Y. Xiaohui, and L. Y. andG. Jiafeng, "Btm: Topic modeling over short texts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2928–2941, 2014.

[18] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," in *36th international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 889–892.

[19] J. Weng, E. Lim, J. J, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *3rd ACM international conference on Web search and data mining (WSDM '10),*, 2010, pp. 261–270.

[20] N. Rajani, K. McArdle, and J. Baldridge, "Extracting topics based on authors, recipients and content in microblogs," in *37th International ACM SIGIR Conference on Research and development in Information Retrieval*, 2014, pp. 1171–1174.

[21] Q. Diao, J. Jiang, F. Zhu, and E. Lim, "Finding bursty topics from microblogs," in *50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2012)*, 2012, p. 536544.

[22] W. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *33rd European conference on Advances in information retrieval*, 2011, pp. 338–349.

[23] H. Sayyadi and M. H. and A. Maykov, "Event detection and tracking in social streams," in *International Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[24] M. Cataldi, L. Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," in *10th International Workshop on Multimedia Data Mining, MDMKDD '10*, 2010, p. 4:14:10.

[25] G. Petkos, S. Papadopoulos, L. Aiello, R. Skraba, and Y. Kompatsiaris, "A soft frequent pattern mining approach for textual topic detection," in *4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, 2014, pp. 25:1–25:10.

[26] F. Abel, Q. Gao, G. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," in *8th Extended Semantic Web Conference (ESWC '11)*, 2011, pp. 375–389.

[27] P. Kapanipathi, F. Orlandi, A. Sheth, and A. Passant, "Personalized filtering of the twitter stream," in *SPIM Workshop at ISWC 2011*, 2011, pp. 6–13.

[28] I. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *AAAI Workshop on Wikipedia and Artificial Intelligence*, 2008, pp. 25–30.

[29] R. Randolf and A. Noack, "Multilevel local search algorithms for modularity clustering," *Experimental Algorithmics (JEA)*, vol. 16, no. 2.3, 2011.

[30] W. Ludo, N. Eck, and E. Noyons, "A unified approach to mapping and clustering of bibliometric networks," *Informetrics*, vol. 4, no. 4, pp. 629–635, 2010.