Trends in Applied Sciences Research

ISSN 1819-3579

Academic
Journals Inc.

www.academicjournals.com

Trends in Applied Sciences Research 9 (5): 219-228, 2014
ISSN 1819-3579 / DOI: 10.3923/tasr.2014.219.228

Concept Mining for Followees Recommendation in Twitter

L.A. Al-Safadi
Department of Information Technology, College of Computer and Information Sciences, King Saud University, Saudi Arabia

ABSTRACT

In this study, author proposed a concept-based recommender of followees in Twitter. The task is performed by mining user interests represented a set of salient concepts in users' tweets. This to support and facilitate the knowledge discovery process from user tweets. The process starts with fetching process of user's last 100 tweets representing his/her latest interest. The system then processes them further to discover those salient concepts of the user. Following that, it identifies those users with similar interest. Similarities between Twitter users are calculated using clustering of users with similar interests. Initial experiments show that the proposed technique is able to perform the task effectively.

Key words: Concept mining, knowledge discovery, followees recommendation, Twitter

INTRODUCTION

Social networking sites have gained considerable popularity in the recent years. This can be explained by the fact that these sites connect people to each other in an easy and timely manner to exchange and share various kinds of information among them.

Twitter gained its popularity through the integration of social networking, blogging and SMS messaging. It allows users to post 140-character text messages, named tweets, to a public timeline of user messages. Users can filter these messages by following the tweets of certain users, named followees. Twitter has grown to accommodate a variety of campaigning. For instance, Twitter was used to provide support to refuges during Jeddah flood in 2010. In business, advertisers and marketers have recognized the potential of Twitter in attracting and maintaining potential customers.

With the rapid expansion of social network users, the content of Twitter is growing exponentially. People are increasingly joining Twitter for networking and discussion because of its compatibility with the different platforms, simplicity and ease of use. Both Java *et al.* (2007) and Krishnamurthy *et al.* (2008) revealed that few users maintain friendship relationship with other users on Twitter. Most of Twitter users behave either as information sources or information seekers. Information seekers follow several users to obtain the information they are interested in. With the growing amount of information, finding the right person to follow is becoming a challenge. On the other hand, to get the most from Twitter, a user must carefully decide in people to follow (followees). Unlike traditional recommendation systems, Twitter has no explicit information available about the user's interests.

Despite the importance and usefulness of Twitter recommendation, limited work has been done in the past. Research use different recommendation strategies for Twitter. Weng *et al.* (2010), Armentano *et al.* (2011) and Yamaguchi *et al.* (2010) used the topology of the social structure in

recommending followees other than the content of the tweets. Chen *et al.* (2010), Phelan *et al.* (2009) and Esparza *et al.* (2010) used Twitter user-generated content to filter information streams, and Guy *et al.* (2009) and Pazzani and Billsus (2007) used user-generated content on the web to extract preferences for recommendation. Hannon *et al.* (2010) used both content and collaborative learning style approaches, based on the followees and followers of users, as well as a number of hybrid strategies.

Chen *et al.* (2010) compared two approaches of followes recommendations; relationship-based and content-based algorithms, finding that the first one is better at finding known contacts whereas the second one is stronger at discovering new friends with common interest.

With this in mind we develop a concept content-based followee recommender for Twitter. We describe the recommender system that is designed to use the user-generated content as a source of recommendation knowledge for followee. The proposed system performs concept mining of user tweets to identify user similarity. Although some work have used user-generated content as a source of profiling information (Guy *et al.*, 2009; Chen *et al.*, 2010; Hannon *et al.*, 2010; Pazzani and Billsus, 2007; Phelan *et al.*, 2009; Esparza *et al.*, 2010), not enough attention has been paid to measure the user's interest. This results in incorrect recommended followees.

TWITTER FOLLOWEES RECOMMENDER HYPOTHESIS

Twitter provides very basic recommenders tool to help people find users to follow. Yet it offers its platform for developers to expand its recommending tool through Twitter API. This provides great opportunity for researchers and developers to develop recommender systems for Twitter.

This study presents a concept content-based recommender system to help people find new users to follow. The basic hypothesizes of our proposed recommender tool are:

- People are attracted to users with similar interests: In Twitter, people find relevant users by browsing their tweets
- Interests are expressed through people exchanged communications (Tweets)
- Frequent concepts in a person's communication are most likely to be the salient concepts that represent user's interest
- Many different concepts may represent the same interest

In Twitter, user's tweets are the source of people interest that can be used as the basis of our recommender system. The following section describes how our proposed recommender system can be used to suggest interesting users to follow.

SYSTEM ARCHITECTURE

The proposed concept-based model involves processing tweets fetched from Twitter. Tweets are provided as inputs to the system. Tweets are processed by identifying interesting concepts. Interesting concepts defined as distinguished concepts that occur frequently in user's tweets.

The architecture of the system in Fig. 1 illustrates the data stores and components of the proposed followee recommender system. The system architecture automatically parses user's tweets and set indexes that represents its interest. The system then finds and returns other users with similar interest.

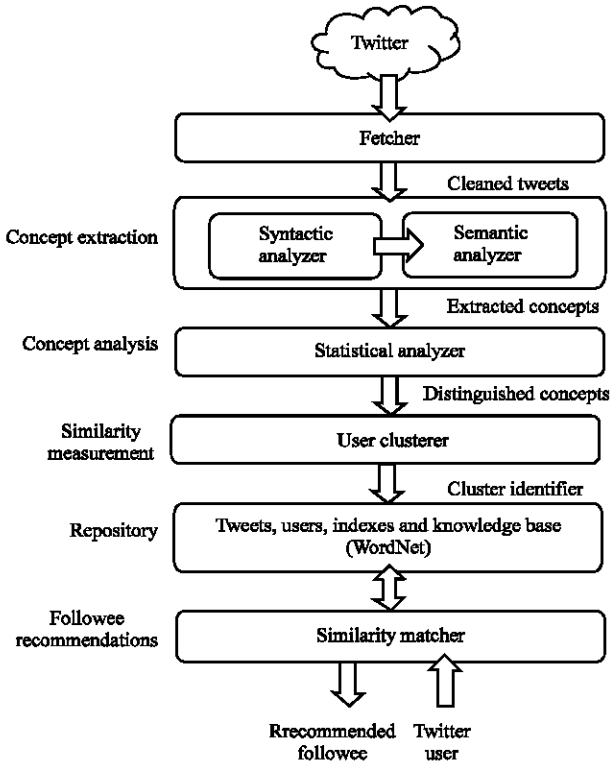

Fig. 1: Proposed system architecture

There are two knowledge bases involved, WordNet and users database. WordNet (Fellbaum, 1998) is a foundation lexical database for the english language to provide shared and common understanding scheme of user tweets.

In addition, the system consists of a number of important components; concept extractor, concept-based statistical analyzer and concept-based similarity measure. The concept extractor is composed of syntactic analyzer and semantic analyzer to process the tweets content and identify salient concepts. The concept-based statistical analyzer weights each term on the tweet, user and Twitter level to determine distinguished concepts. The user cluster groups Twitter users with similar interests. The similarity matcher ranks users within the same cluster with a specific user, which represents user with tweets containing similar interests (interesting concepts).

The system process tweets as follows:

- Fetcher, which fetches the last 100 tweets from Twitter for a specific user and cleans the tweets by removing noise: These include URLs, numbers, dates, usernames, emoticons, positive and negative terms and hash tags. The resulting set of tweets serves as the source for salient concept discovery
- Syntactic analyzer, which analyzes the syntax of the tweet by parsing it and extracting lexical information. Based on the extracted lexical, it locates a trigger phrase for a relationship from each legend and forms from there a triple composed of two phrases and a trigger phrase. Each triple represents a relationship between two concepts

- Semantic analyzer, which semantically analyzes each phrase in each triple by accessing the WordNet and extracting from each phrase its salient concepts
- Concept analyzer, which performs statistical analysis of extracted salient concepts to determine distinguishing concepts
- User clustering, which groups Twitter users based on the similarity of the concepts extracted from their tweets
- Similarity matcher, accesses and ranking user's database in order to get the users with similarity measure higher than the threshold

CONCEPT MINING PROCESS

Concept mining is the ability of computers to understand the meaning of text. It is a relatively new area in machine learning. A word might have multiple meanings and multiple words might be used to represent a particular concept, therefore a word may not carry its exact meaning. This may cause irrelevant content-based followees recommendation in Twitter. This problem is addressed in this study in an attempt to understand the meaning embedded in Twitter content. The objective of our proposed concept mining process is to help the user follow relevant users. Using the proposed technique, the user can quickly find relevant users without browsing through a large number of tweets of irrelevant users.

The proposed concept mining algorithm is based on the work presented by Shehata (2009). The work involves natural language processing as well as statistical analysis towards automatic discovery of patterns in user's tweets. The statistical analysis consists of tweet-based concept analysis, user-based concept analysis, Twitter-based concept analysis and concept-based similarity measure.

Raw text tweets for a particular user are the input to the proposed model and cluster ID in which a user belongs to is the output. The concept mining process can be divided into four main phases; Tweet pre-processing composed of syntactic analysis and semantic analysis of tweets, concept analysis and concept mapping.

Syntactic analysis: For a given user, there may be one or more tweets. Each tweet is usually composed of one sentence. The process of concept extraction and association is applied at the tweet level. Each tweet is parsed and grammatical structures are extracted. Each tweet is made up of a connector word and the two phrases connected by that connector word. The connector word and two phrases together are called a triple.

Prior to information retrieval, a number of preprocessing operations must be performed, these are stopwords removal and word stemming. Stopwords are words that occur too frequently and have little significance for capturing informational meanings, such as "the", "and", "a", ... etc. Stemming algorithms converts any word form into its original base form. One of the popular stemming algorithms is the Porter stemmer (Porter, 1997). In addition, other cleaning operations are performed which include removing URLs, numbers, dates, usernames, emoticons, positive and negative terms and hash tags.

Currently, the system uses a basic parser API called GATE API that is an open source Java implementation. Using the GATE, the part of speech and other syntactic information are analyzed. GATE tokenizer package tokenizes text into words and sentences.

Semantic analysis: After triples are processed, each triple is further analyzed by semantic analysis. This involves looking for concepts in each phrase. One of the difficult problems is the ambiguity of some concepts, which means that the concepts may have multiple meanings and/or may appear in different contexts. This may cause irrelevant recommended followees. Therefore, the semantic analysis aims to determine the meaning of the words in the tweets and identify salient concepts from the tweet. From each phrase, a candidate list of concept phrases from the WordNet is extracted.

WordNet a lexical database for the english language (Fellbaum, 1998). WordNet goal is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language. The words or phrases are considered as concepts if said words can be found in the metathesaurus of the WordNet. As new concepts are identified from the tweets, they are translated into their WordNet preferred terms, which are used to index both tweets and user.

Concept analysis: In the concept-based mining model, labeled terms are considered as concepts. The objective behind the concept-based analysis task is to achieve an accurate analysis of the extracted concepts on the tweet, user and Twitter levels. Frequent concepts are most likely to be the salient concepts that represent user's interest. A concept-based statistical analyzer is used to calculate the frequency and weight of each term on the tweet, user and Twitter level.

Calculating conceptual term frequency: To analyze each concept at the tweet level, a concept-based frequency measure, called the conceptual term frequency ctf is used. The ctf calculations of concept c in tweet t and user u are as follows:

At tweet level (ctf): The ctf is the number of occurrences of concept c in tweet t. A concept c can have many ctf values in different tweets that belong to the same user u. Thus, the ctf value of concept c in user u is calculated by:

$$ctf = \frac{\sum_{n=1}^{sn} ctf_n}{sn} \quad (1)$$

where, sn is the total No. of tweets that contain concept c for user u. Taking the average of the ctf values of concept c for user u tweets measures the overall importance of concept c to the meaning of its tweets for user u. A concept, which has ctf values in most of the tweets for a user, has a major contribution to the meaning of its tweets that leads to discover the interest of user u. Thus, calculating the average of the ctf values measures the overall importance of each concept in representing a user's interest through his/her tweets.

At user level (uf): To analyze each concept at the user level, the concept- based term frequency uf, the number of occurrences of a concept c for a user u, is calculated. The uf is a local measure on the user level.

At Twitter level (tf): To extract concepts that can discriminate between users, the number of users containing concept c, is calculated. The tf is a global measure on the Twitter level. This measure is used to reward the concepts that only appear for small number of users as these concepts can discriminate their users among others.

Calculating conceptual weights: User tweets may contain many unrelated concepts, which some may not be key concepts. Therefore, the following is the concept-based weighting $weight_{stat}$ which is used to discriminate between non-important terms with respect to tweet semantics and terms which hold the concepts that present the meaning of the tweet:

$$Weight_{stati} = ufweight_i + ctfweight_i \quad (2)$$

where, the $ufweight_i$ value presents the weight of concept i for user u at the user-level and the $ctfweight_i$ value presents the weight of the concept i for user u at the tweet-level based on the contribution of concept i to the semantics of the tweets in u. The sum between the two values of $ufweight_i$ and $ctfweight_i$ presents a measure of the contribution of each concept to the meaning of the tweets and to the user interests.

$$ufweight_i = \frac{uf_{ij}}{\sqrt{\sum_{j=1}^{cn}(tf_{ij})_2}} \quad (3)$$

$$ctfweight_i = \frac{ctf_{ij}}{\sqrt{\sum_{j=1}^{cn}(ctf_{ij})_2}} \quad (4)$$

where, uf_{ij} value is normalized by the length of the user vector of the term frequency uf_{ij} for user u and the ctf_{ij} value is normalized by the length of the user vector of the conceptual term frequency ctf_{ij} for user u, where j = 1, 2, .., cn and where cn is the total number of the concepts which has a term frequency value for user u.

Concept-based analysis algorithm: Figure 2 illustrates the proposed concept-based statistical analysis algorithm. The algorithm extracts the significant concepts that represent user u interest in O(m) time, where m is the number of concepts. To do so, weights will be given to each concept. The more significant concept will have more weight. The weights can be calculated as follows:

- t is a tweet
- The total No. of users in Twitter is n
- The ctf_{ij} of each concept c_i in t for each user u_j
- The uf_{ij} of each concept c_i for each user u_j
- The tf_i of each concept c_i
- cn is the total No. of the concepts which has a term frequency value in user u

Twitter user clustering: The syntactic analysis, semantic analysis, concept analysis phases prepares data for mining algorithm. The resulted dataset is called a transaction set. This section we apply data mining techniques to the normalized data (users weighted concepts) to discover the group of users with similar interest.

R is a Twitter
u is a Twitter user
L_{uf} is an empty List user-level (L_{uf} is the top concept list)
L_{ctf} is an empty List tweet-level (L_{ctf} is the top concept list)
L_{stat} is an empty List (L_{stat} is the top concept list)
for each tweet t in u do
    c_i is a new concept in t
    for each concept c_i in t do
        compute uf_i of c_i in u
        compute ctf_i of c_i in t in u
        compute tf_i of c_i in R
        compute $weight_{stati}$ of concept c_i
        add concept c_i to L_{uf}
        add concept c_i to L_{ctf}
        add concept c_i to L_{stat}
    end for
end for
sort L_{uf} descendingly based on max(ufweight)
sort L_{ctf} descendingly based on max(ctfweight)
sort L_{stat} descendingly based on max($weight_{stat}$)
return max(ufweight) from list L_{uf}
return max(ctfweight) from list L_{ctf}
return max($weight_{stat}$) from list L_{stat}

Fig. 2: Algorithm concept-based analysis algorithm

Clustering is a known method for discovering patterns in underlying data. Twitter user clustering aims to automatically divide users into groups based on the similarities of their interest. Interests are represented by distinguished concepts extracted from user's tweets and their weights. Each group (or cluster) consists of users that are similar between themselves (have high intra-cluster similarity) and dissimilar to users of other groups (have low inter-cluster similarity). Clustering Twitter users can be considered as an unsupervised task that attempts to classify users by discovering underlying patterns. Clustering can be considered as the most important unsupervised learning problem. Unsupervised learning means that no need to define the correct cluster (output) into which the user (input) should be mapped to. The clustering datasets are Twitter users, concepts as attributes and $weight_{stat}$ as matrics values.

The aim of Twitter user clustering is to investigate ways of improving the precision or recall of Twitter followee recommender tools. Clustering has been proposed for use in many applications such as; results returned by a search engine in response to a users query and document retrieval.

K-means is a clustering techniques that is commonly used. Although, agglomerative hierarchical clustering is another widely used technique that is perceived to provide more precise results than K-means (Jain and Dubes, 1988), yet it is slower.

K-mean clustering algorithm:

- Select K points as the initial centroids
- Assign all points to the closest centroid
- Re-compute the centroid of each cluster
- Repeat steps 2 and 3 until the centroids do not change

Experiment: In our experiment, a prototype was developed using Java language and GATE. GATE is an open-source, world-leading platform for language technology, which provides support for multiple languages and formats, based on established standards, such as Unicode and XML. In our experiment, GATE was used to extract the concepts found in tweets based on WordNet. Java API for WordNet Searching (JAWS) is an API that was used to provide Java applications with the ability to retrieve data from the WordNet ontology. Twitter API was used to retrieve public Twitter accounts and their tweets.

A number of experiments were carried out to measure the performance of the proposed concept mining approach. We applied retrieval on 10 Twitter users each with 100 tweets retrieved from Twitter using Twitter API to use as test cases. The test cases were used to test the performance of the proposed approaches. The precision rate of results is an important retrieval performance of search systems. In our experiment, we would like to retrieve Twitter users who are similar to others. The test sample contained both 'easy' cases and 'difficult' cases. The 'easy' cases contained hash signs, URLs, numbers and exact matching words. The 'difficult' cases contained slang, synonyms, composite words, plurals, abbreviations and required much pre-processing, accordingly. The experiment went through a number of cleaning phases to increase precision of results. The first cleaning phase was to remove hash tags, URLs, numbers, special characters, stop words and non-english words. The second clearing phase removed plurals and abbreviations. The words left are considered salient words. Salient words with entries in WordNet were replaced with list of retrieved synonyms.

A search result was either similar or non-similar to a tweet. The measure precision, recall and F-score are calculated after clusters are formed. Precision (i, j) = No. of members retrieved and similar to member i in cluster j/No. of members of cluster j. Recall (i, j) = No. of members retrieved and similar to member i in cluster j/No. of members similar to member i. F-score = 2 (Precision*Recall)/(Precision+Recall).

The cleaned test samples provided average results of the overall precision 53%, recall 90% and F-score 67%. In comparison with some of the related work in followee recommendation, (Hannon *et al*., 2010) user-generated content recommender performed a precision rate of an average of 22%. Yet, topology-based followee recommender by Armentano *et al*. (2011) performed 68.15%.

Results show that the concept similarity matching performs better than other content-based recommenders. Most of the failures were related to words not existing in WordNet including slangs and multi-words concepts. More comprehensive sample sets are needed to have a better evaluation of the system.

The system was costly in terms of processing time. Tweets pre-processing and extracting synonym for each concept consumed more than 45% of processing time. Future works would be comparing the prototype performance with other language processing tools and ontologies. In addition, to reduce performance time, the list of synonyms can be identified only when no exact word match was identified.

Experiments also showed that most users use the same word to refer to a concept. Hence, changing concept indexing from tweet-level to user-level is expected to reduce the performance time.

CONCLUSION

The objective of this study is to introduce a concept mining method that aims to understand the meaning of tweets, identify similar tweets and accordingly recommend users to follow. The proposed

approach extracts concepts which capture semantics in tweets. This work explains in detail the proposed concept mining process and its algorithm. This framework utilized the features of both natural language processing and statistical techniques of data mining approaches. It used the benefits of three level similarity measures that are tweet level, user level and Twitter level. An initial prototype was developed to test the proposed algorithm on simple test cases. Experiment results show that the proposed concept mining results 31% improvement in precision values for the process of clustering over (Hannon *et al*., 2010) user-generated content recommender. Most failure related to missing entries in WordNet and multi-words concepts. Tweets contained many slangs with no entries in WordNet. It is believed that results would show higher improvement with vertical applications, specialized ontology and formal text. Further work would be conducting thorough experiments to list the limitations of the proposed process and ways to improve the concept-mining process. In addition, our experiments focus only on finding similarities based on similarly between interesting concepts. Future experiments should take relationships between salient concepts into similarity calculation.

ACKNOWLEDGMENT

This study project was supported by a grant from the "Research Center of the Center for Female Scientific and Medical Colleges", Deanship of Scientific Research, King Saud University.

REFERENCES

Armentano, M.G., D.L. Godoy and A.A. Amandi, 2011. A topology-based approach for followees recommendation in Twitter. Proceedings of the 9th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems, July 16, 2011, Barcelona, Spain.

Chen, J., R. Nairn, L. Nelson, M. Bernstein and E. Chi, 2010. Short and tweet: Experiments on recommending content from information streams. Proceedings of the 28th International Conference on Human Factors in Computing Systems, April 10-15, 2010, Atlanta, Georgia, USA., pp: 1185-1194.

Esparza, S.G., M.P. O'Mahony and B. Smyth, 2010. On the real-time web as a source of recommendation knowledge. Proceedings of the 4th ACM Conference on Recommender Systems, September 26-30, 2010, Barcelona, Spain, pp: 305-308.

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database. 1st Edn., MIT Press, Cambridge, MA., ISBN-10: 026206197X, Pages: 423.

Guy, I., I. Ronen and E. Wilcox, 2009. Do you know?: Recommending people to invite into your social network. Proceedings of the 14th International Conference on Intelligent User Interfaces, February 8-11, 2009, Sanibel Island, Florida, USA., pp: 77-86.

Hannon, J., M. Bennett and B. Smyth, 2010. Recommending twitter users to follow using content and collaborative filtering approaches. Proceedings of the 4th ACM Conference on Recommender Systems, September 26-30, 2010, Barcelona, Spain, pp: 199-206.

Jain, A.K. and R.C. Dubes, 1988. Algorithms for Clustering Data. Prentice Hall Inc., Englewood Cliffs, USA., ISBN: 0-13-022278-X, Pages: 320.

Java, A., X. Song, T. Finin and B. Tseng, 2007. Why we twitter: Understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis, August 12-15, 2007, San Jose, California, USA., pp: 56-65.

Krishnamurthy, B., P. Gill and M. Arlitt, 2008. A few chirps about twitter. Proceedings of the 1st Workshop on Online Social Networks, August 17-22, 2008, Seattle, Washington, USA., pp: 19-24.

Pazzani, M.J. and D. Billsus, 2007. Content-Based Recommendation Systems. In: The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, P., A. Kobsa and W. Nejdl (Eds.). Springer, New York, pp: 325-341.

Phelan, O., K. McCarthy and B. Smyth, 2009. Using Twitter to recommend real-time topical news. Proceedings of the 3rd ACM Conference on Recommender Systems, October 23-25, 2009, New York, USA., pp: 385-388.

Porter, M.F., 1997. An Algorithm for Suffix Stripping. In: Readings in Information Retrieval, Jones, K.S. and P. Willett (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA., ISBN: 1-55860-454-5, pp: 313-316.

Shehata, S., 2009. Concept mining: A conceptual understanding based approach. Ph.D. Thesis, University of Waterloo, Waterloo, Ontario, Canada.

Weng, J., E. Lim, J. Jiang and Q. He, 2010. TwitterRank: Finding topic-sensitive influential twitterers. Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, February 3-6, 2010, New York, USA., pp: 261-270.

Yamaguchi, Y., T. Takahashi, T. Amagasa and H. Kitagawa, 2010. TURank: Twitter user ranking based on user-tweet graph analysis. Proceedings of the 11th International Conference on Web Information Systems Engineering, December 12-14, 2010, Hong Kong, China, pp: 240-253.