Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Keyword Extraction and Topic Finding Approaches in Microblogging Systems: A Survey

Prachi Jain*
Master of Computer Engineering
D. Y. Patil college of Engineering, Akrudi,
Pune, Maharashtra, India

Shanthi K. Guru
Department of Computer Engineering
D. Y. Patil college of Engineering, Akrudi,
Pune, Maharashtra, India

Abstract— Microblogging e.g. Twitter as a new form of online communication in which users talk about their daily lives, publish opinions or share information by short posts, has become one of the most popular social networking services today, which makes it potentially a large information base attracting increasing attention of researchers in the field of knowledge discovery and data mining. In this paper, we conduct a survey about existing research on keyword extraction and topic finding from microblogging services and their applications as friend recommendation, and then address some promising future works. We specifically analyse keyword extraction approaches and five approaches of topic finding.

Keywords— Microblogging, friend recommendation, temporal model, interest drifts.

I. INTRODUCTION

In the current era, People are becoming more communicative through expansion of services and multi-platform applications, i.e., Web 2.0 [1] which establishes social and collaborative backgrounds. They now go beyond personal computing, facilitating collaboration and social interactions [3]. They commonly use various means including Blogs to share the diaries, RSS feeds to follow the latest information of their interest and Computer Mediated Chat (CMC) applications to hold bidirectional communications.

Microblogging is one of the most recent products of CMC, in which users talk about their daily lives, publish opinions or share information by short posts. Microblogging has become a convenient way for Internet surfers and average users to communicate with their friends and family members [2], or to express intimate emotions or feelings. It was first known as Tumble logs on April 12, 2005, and then came into greater use by the year 2006 and 2007, when such services as Tumblr and Twitter arose. According to official statistics, there were 111 microblogging sites internationally in May 2007. Among the most notable microblogging services today are Twitter, Tumblr, Plurk and Chinese Sina Weibo [1], to name a few. As a well-developed and widely-used microblogging service, Twitter has spawned great research interest recently.

Using a microblog also has gradually become a habit for a massive amount of users, which leads to an exponential explosion of information in the virtual microblog society on the Internet, making retrieving and identifying needed microblog or related information extremely difficult. Therefore, more and more microblog services are developing novel engines dedicated to recommending user-specific information.

Therefore, in this paper, we specifically focus on Twitter to study the task of information extraction from microblogging services. Twitter provides its users a strict limit of 140 characters per posting (often called tweet) for broadcasting anything they want. Twitter users can subscribe to other users' tweets by following particular users just like most online social networking services do, such as Facebook and MySpace. However, this follower-and-followee relationship in Twitter requires no reciprocation. That is, the user being followed need not follow back. On receiving a tweet, users can comment on that tweet or retweet (identified by `RT') when they find it of some interest, which empowers a tweet to be visible outside its original one-degree subscribing network. To enhance its freestyle feature, Twitter also predefines a special mark-up vocabulary: `@' followed by a username identifier to address that particular user or to initiate a directed conversation, and `#' followed by a sequence of characters to represent hashtags, which add additional context to tweets and facilitate easy search of tweets that contain similar hashtags [2].

Since its birth in October 2006, Twitter has become one of the most notable social networking and microblogging services today, had attracted more than 540 million registered users by December 4, 2013 [1], generating over 300 million tweets and handling over 1.6 billion search queries per day. The rapidly growing worldwide popularity makes twitter potentially a large information base attracting increasing attention of researchers in the field of knowledge discovery and data mining. Actually, information detection from Twitter has long been a hot research topic in the Web Community recently. However, it is worth mentioning that extracting useful information from Twitter is a complex task, more than simply applying the traditional information extraction technologies that have been proved successful in the Web corpus or other social networking sites to the Twitter context.

Twitter has some distinct characteristics, which make the information extraction process more challenging. For example, unlike web documents or blogs, the postings on Twitter are always short due to the 140-character length limit,

so users won't take too much thinking before making a post. This often leads tweets to be noisy, ungrammatical, and full of abbreviations, symbols and misspellings. As a consequence, traditional NLP tools such as POS taggers or Named Entity Recognizers (NERs) cannot be applied directly to Twitter. Nevertheless, these features also bring many new opportunities to researchers on Twitter. For example, the length-limitation of tweets makes it easier to broadcast a posting, thus in turn making the information contained on the Twitter platform fresher and more real-time. This opens chances of using tweets to predict coming trends or detect ongoing events. Besides, unlike other social networking sites such as Facebook and MySpace, the following network on Twitter is directional rather than reciprocal. In other words, users' requests to subscribe to others do not require the target users' approval. The purpose of our work is to conduct a survey about existing research on information extraction from Twitter, as well as friend recommendation based on information extracted.

II. RELATED WORK

In this section, we review several directions of existing works on user's interest analysis in microblogging systems: (A) Keyword extraction approaches (B) Topic Finding algorithms utilizing social network information.

A. Keyword extraction approaches

In this section, we briefly introduce the basic concepts in keyword extraction including its definition and tasks, as well as typical methods used to extract information from the web. Information Extraction (IE) is an automatic extraction process to generate structured data from a collection of unstructured or semi-structured documents. Unlike information retrieval (IR), which is concerned with how to return relevant documents from a corpus for a given query, information extraction systems generate structured information for post-processing, which is crucial to many applications of data integration and search engines. The input of the IE process can be unstructured documents like free text written in natural language or semi-structured documents such as web pages, which are pervasive on the internet. The result of the IE process is data in a structured form, which can be processed automatically by machines. The extraction of structured data from noisy and unstructured sources is a challenging task. One of the typical tasks in information extraction is named entity extraction, which has become an active and hot research topic over the past decade.

The named entity recognition (NER) problem was originally defined at the Message Understanding Conference 6 (MUC-6) in 1996. The goal of the task is to classify every word in a document as falling into one of eight categories: person, location, organization, date, time, percentage, monetary value, and none-of-the-above [8]. They can be the names 498 International Journal of Software and Informatics, Volume 6, Issue 4 (2012) of people, organizations, geographic locations, times, currencies, and percentage expressions. In general, most previous works can be categorized into three approaches, including rule-based approaches, learning-based approaches and statistical approaches.

1) Rule-Based approaches: In order to recognize types of entities on text documents, rule-based approaches define heuristic rules to identify named entities within documents in a particular domain. These rules are built by experts in the domain to extract information about entities. Most rule-based systems regard the representation of rules so that they obtain the efficiency in matching processes and application of rules for extraction. Therefore, several rule representation formats have evolved over the years. The area of rule-based information extraction (IE) has advanced several rule languages and frameworks [9] for constructing such information extraction programs (called annotators). Since extraction is observed as a sequential operation over text, such rule languages and their executions are Mostly based on the theory of grammars and finite state automata [5].

Early entity recognition systems primarily adopted rule-based approaches. They are efficient for domains where there is certain formalism in the construction of terminology. A typical example is the biology domain, where certain types of entities can be extracted by domain-specific rules with sufficient accuracy. Also it has been successfully applied in open information extraction where information redundancy is available for relatively simple types of entities [10].

One of the advantages of this approach is that execution time of rule-based systems is shorter than other methods. In addition, developers can easily control the rules to obtain certain optimization for some specific domains, such as the extraction of phone numbers, zip codes, dates, and time. However, this approach requires experts to define the rules for extraction, which can be rigid and not general enough to cover all cases in real data. This may directly affects the completeness of entity types in the results of rule-based systems. As a result, the effort of research has moved towards more robust learning based approaches since they have been introduced.

2) Learning-Based Approaches: Machine learning is a way to automatically learn to identify complex patterns or Sequence labelling algorithms and make intelligent conclusions based on data. Learning algorithms are methods able to consume features of training data to automatically induce patterns for recognizing alike information from unseen data. Learning algorithms can be generally classified into three types: supervised learning, semi-supervised learning and unsupervised learning. Supervised learning utilizes only the labelled data to generate a model. Semi-supervised learning aims to combine both the labelled data as well as useful evidence from the unlabelled data in learning. Unsupervised learning is designed to be able to learn without or with very few labelled data.

Supervised methods are the algorithm that acquire a model by looking at annotated training examples. Among the supervised learning algorithms for NER, considerable work has been done using Hidden Markov Model (HMM) [4], Decision Trees [15], Maximum Entropy Models (ME) [8] and Conditional Random Fields(CRF) [11]. Typically, supervised methods either learn disambiguation rules based on discriminative features or try to acquire the parameter of presumed distribution that maximizes the likelihood of training data.

The term *semi-supervised* is relatively recent. The main method for semi-supervised learning is called "bootstrapping" and comprises of a small degree of supervision, such as a set of seeds, for starting the learning process. Semi supervised learning algorithms [13] use both labelled as well as unlabelled corpus to create their own assumptions. Algorithms usually start with small amount of seed data set and create more hypothesis or assumptions using large amount of unlabelled corpus. It make use of unlabelled data for training typically a small quantity of labelled data with a large amount of unlabelled data [14]. The semi-supervised algorithm is used to overcome the problem of insufficient annotated corpus and data sparsity problem. Semi-supervised usually starts with small volume of annotated corpus, large volume of unannotated corpus and a minor set initial hypothesis or classifiers. With each iteration, further annotations are generated and stored until a definite threshold occurs to stop the iterations [8]. A most important problem with supervised learning is requirement of large number of features. For learning a good model, a strong set of features and huge annotated corpus is needed. Many languages do not have large annotated corpus presented at their disposal. To deal with insufficient annotated text across domains and languages, the next unsupervised techniques for NER have been proposed.

The classic approach in *unsupervised learning* is clustering. The example is to collect named entities from clustered groups based on the similarity of context. There are some other unsupervised methods too. Mostly, the techniques depend on lexical resources on lexical patterns and on statistics computed on a large unannotated corpus [12].

3) Statistical Approaches: The underlining idea of the statistical approach for NER is to solve the problem of entity recognition by two phases, including decomposition of unstructured texts and a phase of labelling the parts of decomposition. The parts of decomposition are commonly represented in one of two prevalent forms: tokens and word chunks. In the labeling phase, a model which is firstly trained from a training dataset is used to identify information of entities from unstructured texts. One of the ways to assign labels for tokens is to view the problem of token labeling as the problem of classification in which the model must determine whether a token is assigned a particular label or not. Therefore, any existing classifier can be used to classify tokens. Reference is an example of research work that used a Support Vector Machine (SVM) [16] to extract meta-data of citations.

III. TOPIC FINDING APPROACHES

To discover the user interest, only keyword extraction is not sufficient. As the persistence of synonymy, it is required to find the unseen topics from the keyword usage patterns. As the goal is to find topics that each microblogging user is interested in rather than topics that each microblog is about [1].

A. LDA Model

The general idea of Latent Dirichlet Allocation (LDA) is based on the hypothesis that a person writing a document has definite topics in mind [18]. This model discovers users' topic distributions according to their keyword usage patterns [1]. LDA is an unsupervised, reproductive model that proposes a stochastic procedure through which words in the documents are produced by finding latent semantic topics in huge collections of text documents. The key insight into LDA is the source that words contain strong semantic data about the document. Hence, it is reasonable to undertake that documents on roughly like topics will use the similar group of words. Latent topics are therefore discovered by recognizing groups of words in the corpus that repeatedly occur together within documents. Learning is unsupervised in this model because the input data is partial: the corpus make available only the words within documents; there is not any training set with topic or subject annotations [17].

B. In-Degree

Twitter is one of the most prominent micro-blogging services, employs a social-networking model called "following", in which every twitterer is permitted to choose whom he/she wants to follow without looking for any permission. On the other hand, he/she may also be followed by others deprived of granting permission first. In one instance of "following" link, the twitterer whose posts are being followed is called the "friend", whereas the one who is following is titled the "follower". Since users seem to follow influential people on the microblogging platform [6], it is good to add such people as users' friends. In-degree method is used to identify the influential twitterers. Twitterers follow others because they are interested in the topics the other friend share in tweets and the friend follows back because they find sharing of similar topic interest. TwitterRank [6], approach use the in-degree method to measure the influence of twitterers or the number of followers and then suggests friends based on users' influence.

C. User-Based

In user-based approach [20], first the similar users are identified. The similar user can nearest neighbour and which can be identify using vector space model or parson correlation. Each user is treated as vector in n-dimensional space and similarity between two vectors i.e. active user and other user are computed.

This method is based on assumption that users who have more common friends the chances of having similar to each other is more. So the recommendation of friend is done based on number of common friends of user's. The process of calculating similarity or similar users is crucial for the accuracy of prediction because the prediction of the unknown value depends on the corresponding values of similar users [7].

D. Online-LDA

Latent Dirichlet Allocation (LDA) [1] covers the generative model to attain the capacity of generalizing the topic distributions so that the model can be used to generate unobserved documents as well. Online-LDA is the online version

of LDA which automatically captures the thematic patterns and discover topics of text streams over time. Online-LDA approach allows LDA to work in online fashion which creates the up-to-date model which is the mixture of topics per document and mixture of words per topic when a new or set of documents appears [19].

To allow LDA to act as on-line version of LDA on data streams, Online-LDA model deliberates the temporal related information and consider that the documents are distributed in time slices. To each time slice, a topic model along with K components is used to model the newly attained documents. The generated model, at any point of time, is used as a prior for LDA at the successive time slice, when a fresh data stream is accessible for processing. The Online-LDA discover the interesting patterns by just evaluating a fraction of data at a time [19]. The Online-LDA is sometimes equal to, and sometimes better than, the original LDA in guessing the likelihood of hidden documents.

IV. DISCUSSIONS

Finally, we summarize below three promising directions on the topic of information extraction from microblogs, which we find through this survey. We hope it can help new comers to get started in this field quickly.

1) Keyword extraction from microblogs: There are three types of sources in microblogs from which we can mine useful information, i.e., the metadata contained in user profiles or tweets (e.g. interests, locations, timestamps, etc.), the content of tweets, and the network structure of following, mentioning and retweeting.

2) Topic finding from microblogs: Latent Dirichlet allocation (LDA) achieves the capacity of generalizing the topic distributions so that the model can be used to discover users' preferences. Users' interests are not static, their interests may change as time goes by. So, LDA calculates users' potential interests on others according to user similarities over different periods of time via a temporal function called T-LDA, where T represents temporal function. The in-degree model perform better than the user-based approach. Compared with the three LDA-based models, T-LDA performs better than LDA and OLDA. Adding temporal influence may differentiate users' current interests from the past interest. In T-LDA model, the exponential decay function increasingly discounts the history of past behavior, which focus current interests along with long-term interests. Adding interest drifts in LDA is good indication of user's preference and makes LDA as T-LDA.

3) Where to use the information extracted from microblogs: In fact, the extracted information can be utilized in many applications which facilitate users' daily life, such as online political disclosure prediction, crisis detection and management, user clustering and community detection, user ranking and friend recommendation, personalized information service, and regional targeted advertising, to name a few. These topics have covered most of the leading works during the recent five years, and are still vibrant in the microblogging research community, with new services and applications coming into being every day.

V. CONCLUSIONS

This paper has reviewed the recent state of the art in the literature of information extraction and topic finding from microblogging services (exemplified with Twitter). We specifically focus on different types of methods used for keyword extraction and discussed topic finding approaches used in microblogging for friend recommendation, and then propose some suggestions for future work. In our opinion, this paper can serve as guidance to researchers interested in this field.

ACKNOWLEDGMENT

The authors would like to thank the publishers, researchers for making their resources available and teachers for their guidance. We also thank the college authority for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

REFERENCES

[1] Zheng, Nanning, Seunghyun Song, and Huihui Bao. "A Temporal-Topic Model for Friend Recommendations in Chinese Microblogging Systems."

[2] Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling. "Characterizing Microblogs with Topic Models." ICWSM 10 (2010): 1-1.

[3] Wang, Fei-Yue, et al. "Social computing: From social informatics to social intelligence." Intelligent Systems, IEEE 22.2 (2007): 79-83.

[4] Bikel, Daniel M., Richard Schwartz, and Ralph M. Weischedel. "An algorithm that learns what's in a name." Machine learning 34.1-3 (1999): 211-231.

[5] Reiss, Frederick, et al. "An algebraic approach to rule-based information extraction." Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. IEEE, 2008.

[6] Weng, Jianshu, et al. "Twitterrank: finding topic-sensitive influential twitterers."Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010.

[7] Wu, Jian, et al. "Predicting quality of service for selection by neighborhood-based collaborative filtering." Systems, Man, and Cybernetics: Systems, IEEE Transactions on 43.2 (2013): 428-439.

[8] Borthwick, Andrew. A maximum entropy approach to named entity recognition. Diss. New York University, 1999.

[9] D. Ferrucci and A. Lally, UIMLA: An architectural approach to unstructured information pro cessing in the corp orate research environment,"Nat. Lang. Eng., 2004.

[10] Chiticariu, Laura, et al. \Domain adaptation of rule-based annotators for named-entity recognition tasks." Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010.

[11] Duan, Huanzhong, and Yan Zheng. \A study on features of the CRFs-based Chinese Named Entity Recognition." International Journal of Advanced Intelligence 3.2 (2011): 287-294.

[12] Keretna, Sara, et al. \Classification ensemble to improve medical Named Entity Recognition." Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on. IEEE, 2014.

[13] \Semi-supervised learning." Wikipedia: The Free Encyclopedia.Wikimedia Foundation, Inc. 2nd March 2015. Web.21 April 2015.hhttp://en.wikipedia.org/wiki/Semi-supervised learning

[14] Irmak, Utku, and Reiner Kraft. \A scalable machine-learning approach for semi-structured named entity recognition." Proceedings of the 19th international conference on World Wide Web. ACM, 2010.

[15] Isozaki, Hideki. \Japanese named entity recognition based on a simple rule generator and decision tree learning." Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2001

[16] Isozaki, Hideki, and Hideto Kazawa. \Efficient support vector classifiers for named entity recognition." Proceedings of the 19th international conference on Computational linguistics-Volume1.Association for Computational Linguistics, 2002.

[17] Hu, Diane J. "Latent dirichlet allocation for text, images, and music." University of California, San Diego. Retrieved April 26 (2009): 2013.

[18] Krestel, Ralf, and Peter Fankhauser. "Tag recommendation using probabilistic topic models." ECML PKDD Discovery Challenge 2009 (2009): 131.

[19] AlSumait, Loulwah, Daniel Barbará, and Carlotta Domeniconi. "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking." Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008.

[20] Zhao, Zhi-Dan, and Ming-Sheng Shang. "User-based collaborative-filtering recommendation algorithms on hadoop." Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on. IEEE, 2010.

