Int J Software Informatics, Volume 6, Issue 4 (2012), pp. 495–522
International Journal of Software and Informatics, ISSN 1673-7288
©2012 by ISCAS. All rights reserved.

E-mail: ijsi@iscas.ac.cn
http://www.ijsi.org
Tel: +86-10-62661040

Information Extraction From Microblogs: A Survey

Wen Hua[1], Dat T. Huynh[2], Saeid Hosseini[2], Jiaheng Lu[1], and Xiaofang Zhou[1,2]

[1](School of Information, Renmin University of China, Beijing 100872, China)
[2](School of Information Technology and Electrical Engineering, University of Queensland, Australia)

Abstract Microblogging(e.g. Twitter, http://twitter.com), 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 information extraction from microblogging services and their applications, and then address some promising future works. We specifically analyze three types of information: personal, social and travel information.

Key words: microblogging; information extraction; twitter

Hua W, Huynh DT, Hosseini S, Lu JH, Zhou XF. Information extraction from microblogs: a survey. ***Int J Software Informatics*****, Vol.6, No.4 (2012): 495–522. http://www.ijsi.org/1673-7288/6/i142.htm**

1 Introduction

In the current era, People are becoming more communicative through expansion of services and multi-platform applications, i.e., the so called Web 2.0 which establishes social and collaborative backgrounds. 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. It was first known as Tumblelogs 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[97]. Among the most notable microblogging services today are Twitter, Tumblr, Plurk and Chinese Sina Weibo, to name a few. As a well-developed and widely-used microblogging service, Twitter has spawned great research interest recently. 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

This research is partially supported by 863 National High-Tech Research Plan of China under Grant No. 2012AA011001.

Corresponding author: Xiaofang Zhou, Email: zxf@itee.uq.edu.au

Received 2012-02-28; Accepted 2012-08-06.

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 markup 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.

Since its birth in October 2006, Twitter has become one of the most notable social networking and microblogging services today, "with over 300 million users as of 2011, generating over 300 million tweets and handling over 1.6 billion search queries per day"[98]. 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 realtime. 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. Then in this situation, how to guarantee privacy has also become a promising while challenging research topic.

The purpose of our work is to conduct a survey about existing research on information extraction from Twitter, as well as its applications in real life, and recommend some promising future works to researchers interested in this field. We specifically analyze three types of information:

- `Personal Information` – information that is generally contained in a user's profile, including demographic features such as age, gender, ethnicity and home address; and other features including health status, political orientation, business affinity, user interest and so on.

- `Social Information` – information that identifies users' relationships or interactions with others, including topological structure, community distribution and even detailed social relationships such as co-workers, family members and so on.
- `Travel Information` – information about users' current location, travel history, or whether or not a specific user would be on vacation away from his house and its timing.

The remainder of this paper is organized as follows: Section 2 presents a brief introduction to the basic concepts and widely-used technologies in information extraction; Section 3 explains opportunities as well as challenges in applying traditional information extraction technologies to microblogs; Section 4, 5 and 6 detail related research works on personal information extraction, social information extraction and travel information extraction separately; Then we propose a summary of existing research and outline the avenues of future work in Section 7; The conclusion is addressed in Section 8.

2 Information Extraction Basis

In this section, we briefly introduce the basic concepts in information 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. Named entities are those that are referred to by names such as people, organizations, or locations[32]. According to the study and analysis of Guo et al.[34], named entities occur in about 71% of search queries. Nevertheless, current search engines such as Google and Bing, which support users to search information on the web according to their queries, are mainly based on keyword or text matching techniques and do not capture the semantic information of objects and relations between them. Therefore, the problem of entity extraction, which includes identifying named entities, their attributes as well as the relations between entities, is an indispensable task not only in query answering but also in knowledge discovery from the web environment.

2.1 Named entity recognition

The named entity recognition(NER) problem was originally defined at the Message Understanding Conference 6 (MUC-6) in 1996[32]. The goal of the task is to identify names and types of entities from text documents. They can be the names of people, organizations, geographic locations, times, currencies, and percentage expressions. In general, most previous works can be categorized into two approaches, including rule-based approaches and statistical approaches.

2.1.1 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. For example, the Common Pattern Specification Language in FASTUS[44], regular expression in WHISK[88], JAPE language in GATE[20], Datalog expressions in DBLife[85], and algebraic language in Avatar[77].

One of the advantages of this approach is that execution time of rule-based systems is shorter than other methods[85,77]. 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 affect the completeness of entity types in the results of rule-based systems.

2.1.2 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 labeling 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 [38] is an example of research work that used a Support Vector Machine (SVM) to extract meta-data of citations.

Nevertheless, the labels of adjacent tokens are seldom independent of each other, and can be used to determine the label of a token. Consequently, different models were proposed to capture the dependency between the labels of adjacent words, such as Hidden Markov Models (HMMs)[6], Maximum entropy Markov models (MEMM)[61], and Conditional Random Fields (CRFs)[56]. Currently, CRF based methods are the current state-of-the-art and outperform all previous machine learning based methods in both theory and experimental evaluations for the problem of sequence labeling in the machine learning based approach[68,84].

Like other learning methods, the training dataset plays an important role in training the extraction model. The performance of training models are directly affected by the quality and quantity of training data. Therefore, the learning methods in this approach requires a lot of time and laborious work to manually build such a training dataset. This limits the usage of these methods in web-scale applications.

2.2 Relation extraction

Relation extraction is one of the sub-tasks of entity extraction. It is motivated by the requirements of extracting and structuring the attributes of named entities from documents so that machines can process automatically. The goal of this task is to identify and extract semantic relations between entities within text. Each relation has a type signature that decides the entity types in the arguments of that relation. For example, the relation birthPlace $\subseteq$ Person × Location, graduatedAt $\subseteq$ Person × University are binary relations.

Most previous works concentrated on the extraction of binary relations from documents[2−4,14,51,]. Meanwhile, some others look to the extraction of higher-arity relations or records from web[11,12,23,33,35,58,92]. In general, previous works modeled relation extraction problems in one of the two following scenarios. The first one is to design algorithms that identify the type of relationship between given entity pairs in a document. The second scenario in relation extraction is to retrieve all entities that satisfy a given relationship type. Based on the scenario that the relation extraction problem is modeled, one of the following approaches: a supervised approach, a semi-supervised approach, or an unsupervised approach is employed to extract relations between entities on documents.

2.2.1 Supervised approaches

The supervised learning approach is a technique for identifying a relation type between given entity pairs in sentences or textual documents. Given a pair of entities in a sentence, in order to determine which relation is mentioned in the sentence, the supervised learning approach exploits techniques in natural language processing (NLP) to parse the sentences into well-defined structures. Due to this process, various syntactic and semantic properties in the well-defined structures of a sentence are employed as a set of features for the detection of relations between entities.

This approach supposes that there is a list of entities occurring in a document, and its goal is to identify all occurrences of the relation instances in a fixed set of relation types in that document. In relation extraction from a natural language text, it is assumed that two entities in a relation are in close proximity to each other, or occur in the same sentence. Therefore, the relation extraction problem can be formulated as the problem of identifying whether there are any relationships between two particular entities in a sentence, or not. Most previous work defines a kernel function based on the structured representation of a sentence and combined with Support Vector Machine (SVM) to extract relations from text[9,65,67,75,76]. This kernel function can be considered as a similarity measure between the structures that describe entity pairs in a particular relationship. It is studied from a labeled training dataset and then used to determine the label of a novel entity pair by SVM algorithm. In other words, the problem of relation extraction is formulated as a classification problem and relations of unseen entity pairs are classified by a target function which was learnt from a manually labeled training dataset.

In general, this learning approach for relation extraction requires expensive, deep linguistic parsing techniques in NLP. Moreover, a training dataset of sentences with correct labels must be manually built to train the model which is then used to predict new semantic relations. This makes supervised methods difficult to extend to several types of relations and apply to large-scale applications for extracting information from the web. In addition, this approach requires POS tagging and sentence parsing which may consume resources and is prone to errors. Syntactical parsing may deal with the problem of syntactic ambiguity when there are more than one syntax tree representing a sentence. Therefore, in spite of extensive research in the NLP community, the current accuracy values of this approach for relation extraction still range in the neighborhood of 50%–75%, even in the ACE standard benchmark dataset[76].

2.2.2 Pattern-Based approaches

An alternative scenario of extracting relationships between entities is to extract all entity pairs of one or more given relationships occurring in a document or a corpus. An obvious idea to solve this problem is to exploit the representation of those relations on documents and define patterns or rules to extract information of entities.

The usage of patterns for information extraction from natural language documents has a long history. In 1992, Hearst et.al[40] proposed the usage of syntactic patterns to identify instances of predefined relationship types from free text. These patterns are defined by using regular expressions on Part of Speech (POS) of words in sentences. For example, the patterns of the form "<Noun> such as <List of Noun Phrases>" was used in Ref. [40] to extract *instanceOf* relations, which is also called hypernym relations.

Although Hearst patterns can yield relatively high precision, the patterns which are manually defined become difficult to adapt so as to deal with arbitrary target relations. Therefore, Brin[8] suggested a new method, called DIPRE, which exploited the duality of facts and patterns for extracting relation instances between entity types of authors and books. In Brin's idea, each relation r is described by specifying the types of entity pairs that form the arguments of relation r. An initial set of seed facts for one or more relations is firstly utilized to find automatically the markup, textual, or linguistic patterns of relations from a corpus. These patterns are then applied to identify new fact candidates from the corpus. Reference [78] is a similar work also using this method to extract information in the terrorism domain. The way of studying such patterns from a set of seed entities is also called the bootstrapped method or semi-supervised learning method in the literature. These patterns were improved, enriched, and deployed in several systems, such as Snowball[2], StatSnowball[104], KnowITAll[24,25], TextRunner[3,101], LEILA[89], and SEAL[93,94].

In general, an assumption in the bootstrapped method is that any given entity pairs in an initial seed set cannot participate in more than one relationship with each other. This may not be difficult to obtain in practice because not all sentences containing an entity pair support the relationship type. For example, in the two following sentences: "Tom is a friend of Jack" and "Tom is a colleague of Jack", the Tom and Jack entities participate in two different relationship types. In addition, this approach requires that all entities have been marked in a document. This may lead to some difficulties when entities in the initial seed set occur in different aliases in text. Moreover, bootstrap based methods demand a good evaluation measure and strategy to assess and control the quality of generated patterns.

2.2.3 Unsupervised approaches

In general unsupervised approaches are domain-independent and target different kinds of resources on the Web, including texts, HTML tables, and lists. One method, for the unsupervised approach to extract relations of entities on web without using human knowledge, is to employ the web environment as a corpus to cluster the pairs of co-occurring named entities according to the similarity of the context words between them. This idea was mentioned in Ref. [43] which proposed unsupervised clustering techniques to cluster the noun phrases occurring as subject or object of a given verbal phrase. For example, the word "wine" may occur with "drunk" or "produced", but not with "eaten" or "driven". Reference [86] was another work which clustered all possible relations from text and represented them in tables. In these clustering-based approaches, relation labels will be manually assigned for each cluster of pairs of named entities. Therefore, it is difficult to apply these techniques to large-scale extraction systems. In order to overcome this limitation, Ref. [7] have recently proposed a clustering algorithm to extract relations between entities from unlabeled data. This work clustered simultaneously the entity pairs and vocabularies in different sentences. Due to this process, it was able to identify representative lexical patterns of semantic relations and use them to propose the relation names for entity pairs. However, relations extracted by the system are not in well-defined representation because they are generated from word phrases between entities in documents.

Besides the clustering based techniques mentioned above, some recent research studies proposed novel methods to extract tables or records from the web environment. According to the study of[12], there are over 100 million tables on the web. The meaning of these tables, however, is rarely explicit from their data. Therefore, in Refs. [11,12,13], authors described the WebTables system which exploited tables on the web as a source of high quality relational data for search engines. The statistics of the co-occurrences of attributes in tables could be used to implement a column thesaurus and to propose column auto-completion in queries. Inspired by the benefits of web tables and in order to furnish more semantics for tables on the web, some recent works have extracted information from web table. For example, Limaye et al.[58] proposed a graphical model for annotating table columns on the Web with labels of types, binary relations, and table cells with entity identifiers from an ontology, such as YAGO[90]. Meanwhile, Ref. [92] has recently developed a statistical reasoning model to determine a label for each column and binary relationships in tables. Although several efforts have been attempted in the literature to extract information of entities, existing information extraction approaches are not sufficient to provide efficient solutions to the problem because of the variety of noise in the content of web pages.

3 Information Extraction From Microblogs

The emergence of microblogging services is changing the form people share information on the web. People often access a social network such as Tumblr or Twitter to retrieve news, videos, or comments from their friends. In such a system, a large amount of posts or tweets are posted every day. According to a report on the Tumblr staff blog[1] in 2010, there were average two million new posts and 15,000 new users

[1] `http://staff.tumblr.com/post/434982975/a-billion-hits`

every day. Moreover, a report from Twitter[2] in June 2010 said that it handled about 65 million tweets per day.

Due to the large volume of data available on microblogging sites, it is natural to consider the automatic methods of information extraction to capture semantic meanings of entities in their data for processing. For example, in order to extract the abovementioned three types of information covered in our work(i.e., personal, social and travel information), the most straightforward way is to treat personal information as attributes of entities, social information as relations between entities, travel information as lists of location entities, and then directly apply traditional information extraction technologies to the collection of microblogs or tweets. However, this attempt has been proved to be difficult and unsuccessful because of the following reasons:

- `Ungrammatical Sentence.` Unlike documents on the web, tweets on microblogging services are always length-limited. For example, Twitter allows users to post only 140-character messages. This length-limitation often leads tweets to be noisy and ungrammatical, which makes traditional NLP tools such as POS tagger inappropriate to use.

- `Informal Writing.` Tweets often contain noisy texts such as abbreviations, symbols and misspellings, which consequently brings great difficulties to analyzing the content and meanings of tweets. The sentence "I don't knw wht s gona happen" is an example of a message which people briefly write in abbreviated form on a microblogging site.

Some research studies have been recently proposed to deal with these two problems. For example, Ref. [37] proposed methods to normalize texts in tweets, which showed that the majority of ill-formed words in Twitter are based on morphophonemic variations. Although they often miss some letters or have extraneous letters, the target words for ill-formed words can be effectively extracted. In Ref. [37], a list of candidate canonical lexical forms is generated for each noisy word, then the similarities between the candidates and ill-formed words are calculated before they are ranked to choose the best candidate for each noisy word. Moreover, the study of Ref. [31] showed that different populations of users used different types of lexical transformations in Twitter microtexts. The results of these methods are useful for the preprocessing step of the upstream tasks such as POS tagging and NER. Similarly, Ref. [27] was a research work which built a POS tagger for tweets using 20 coarse-grained tags.

For the problem of entity extraction, Ref. [26] investigated the use of Amazons Mechanical Turk[3] and CrowdFlower[4] to annotate named entities in tweets and train a CRF model to evaluate the effectiveness of human labeling. In Ref. [59], the authors combined a k-nearest neighbor (KNN) classifier with a CRF model to leverage cross tweets information, and adopted a semi-supervised learning to deal with unlabeled tweets. Moreover, Ref. [79] was a recent study which applied the Labeled LDA topic

[2] `http://blog.twitter.com/2010/06/big-goals-big-game-big-records.html`

[3] `https://www.mturk.com/mturk/`

[4] `http://crowdflower.com`

model[73] and exploited Freebase dictionaries as a source of distant supervision for NER problem on Twitter microblogs. Reference [4] was another work which proposed a graphical model, a modification of CRF, to extract relations of artists and venues associated with musical performances in Twitter.

The studies referred to above all focus on the content of tweets as a source for information extraction, and try to solve the problem of tweets being noisy and ungrammatical. However, as information platforms and social networking sites at the same time, microblogging services also provide researchers with many other sources besides tweets, such as profile data and topology data, which actually promises more opportunities for information extraction from microblogs. In the following three sections, we separately investigate existing works on the topic of personal, social, and travel information extraction from microblogs, as well as their applications in the real world.

4 Personal Information Extraction

Personal information can be defined as information generally contained in a user's profile, including demographic features such as age, gender, ethnicity and home address; and other features including health status, political orientation, business affinity, user interest and so on. Twitter provides its users with the mechanism to edit their profiles with the following free-text fields:

- Screen Name(e.g. kgarnette21, queenofen)
- Full Name(e.g. Kevin Garnette, Queen of England)
- Picture(e.g. photo identifying user's face)
- Location(e.g. Brisbane, Somewhere on the earth)
- URL(e.g. user's website, Facebook page)
- Biography(e.g. family member, education status)

Note that all of the above fields, except Screen Name, are completely optional, and even if users do input these information, some of them might be inaccurate or nonsensical(e.g. "Somewhere on the earth" in the "Location" field). Besides, many other features such as age, gender, ethnicity, political orientation and user interest are altogether missing, while they might be of tremendous use in such applications as demographic analysis of Twitter usage, Location Based Services and User Recommendation, just to name a few.

In this section, we first summarize basic approaches used in the area of personal information extraction from Twitter, then focus on User Interest Mining which has attracted massive attention recently, followed by a brief introduction of applications where the extracted personal information can be utilized.

4.1 Basic approaches

According to the information sources used, we specifically divide existing research studies into four categories: Profile-Based Approach, Content-Based Approach, Network-Based Approach and Hybrid Approach, which we detail below.

4.1.1 Profile-Based approaches

B. Heil and M. Piskorski[42] inferred users' gender by cross-referencing their "Full Name"s in profiles against a database of 40,000 strongly gendered names. A similar approach was applied by M. Cheong and V. Lee[18], who utilized the United States Census statistics on 5,494 ethnically-diverse first names, and matched them against users' given or middle names in the "Full Name" fields, thus probabilistically determining users' gender at an accuracy of 86.6%. In another work, M. Cheong and V. Lee[17] also used "Full Name"s, as well as profile images to identify users' gender manually. M. Pennacchiotti and A.-M. Popescu[69] in their work tried to apply regular expression matching on users' "Biography" field to infer age(e.g. $(I|i)(m|am|'m)[0-9]+(yo|yearold)$), gender and ethnicity (e.g. $white(man|woman|boy|girl)$), but the results turned out to be of low accuracy. They also investigated the use of profile pictures to infer users' gender and ethnicity, but still found it misleading, as users often fill the "Picture" field with photos of a celebrity or other images instead of representing themselves.

4.1.1 Content-Based approaches

L. Humphreys et al.[48] found that "personally identifiable information(such as email addresses and phone numbers) are rarely detected in tweets, but a quarter of tweets do include information regarding how people feel or when people are engaging in activities and where they are". Thus, Content-Based Approaches are mainly used to extract such information as home address, health status, political orientation and user interest from tweets. For example, H. Mao et al. Reference [60] applied regular expression matching on a user's tweets to automatically detect whether or not he had a certain disease(exemplified with cancer in their work). The simple classification rule they used is "If has/have/had. . . cancer exists and (dog/cat/kitty/puppy and don't/doesn't have/has/had) doesn't exist in a tweet, it is sensitive; otherwise, it is not.", which indicated a 76% precision in a quantitative experiment.

Despite the classification accuracy achieved in H. Mao et al.'s work, regular expression matching is obviously too simple a way for content analysis of tweets. Instead, Z. Cheng et al.[16] employed maximum likelihood estimation to calculate the probabilistic distribution over cities for each word in the whole Twitter vocabulary, then determined a user's home address by aggregating across all words in that particular user's tweet vocabulary. Furthermore, they found that some words or phrases are local, i.e. they are mostly used in a small geographical area. For instance, "howdy" is a typical greeting word in Texas, thus indicating that the user who often tweets with "howdy" is most likely to live in or near Texas. Therefore, they incorporated a classification component to automatically identify nonlocal words, which were removed from a user's tweet vocabulary when aggregating the city distribution of words. This idea of taking into consideration meaningful words and phrases as a feature vector was also adopted by D. Rao et al.[74] who built binary classifiers using Support Vector Machine(SVM) to identify user's gender, age, regional origin and political orientation. They quantitatively compared two classification models: 1) the Sociolinguistic-feature model, in which they manually constructed a list of sociolinguistic features, such as simleys, repeated alphabets; 2) the Ngram-feature model, in which unigrams and bigrams of the tweet text were derived and weighted by normalized term frequency.

Their experiment showed different comparing results over gender, age, regional origin and political orientation, thus indicating diverse language usage across these four fields in modern informal communication.

4.1.3 Network-Based approaches

P. T. Metaxas and E. Mustafaraj[62] applied graph-theoretic techniques to users' political orientation prediction. They drew the following network using a force-directed algorithm which draws nodes sharing many neighbors closer than those who do not, and detected two user groups representing Liberals and Conservatives respectively. Their 98% precision in predicting the top 200 users' political orientation in their dataset simply by observing their following behaviors revealed that users tend to follow similar users. Inspired by this result, J. Golbeck and D. L. Hansen[28] also adopted a Network-Based Approach to understand Twitter media bias. They first applied liberal/conservative scores obtained from Americans for Democratic Action(ADA) to Congress people using Twitter, then mapped the scores onto their followers by a simple average function, and finally mapped the inferred scores of followers onto the Twitter accounts of media outlets also by an average function. Their predictions of media bias were similar to the liberal/conservative leanings as presented in their prior work[22], which inferred political orientation of webpages and their associated news outlets based on co-citation of hyperlinks.

4.1.4 Hybrid approaches

E. Mustafaraj and P. T. Metaxas[66] tried to discover a user's political orientation exposed in his edited retweets by applying a co-training algorithm to two independent sources: a textual source(the text of the edited retweets) and a structural-behavioral source(the social network of the user and his retweeting habit). J. D. Burger et al.[10] in their work into discriminating the gender of Twitter users, employed both the content of tweet text and three fields from the user profile(i.e. Screen Name, Full Name and Biography). They constructed both word-level and character-level ngrams from each of these four fields, and expressed each classification feature as a simple Boolean indicator representing presence or absence of the corresponding ngram, and then applied the Balanced Window2 algorithm to train this classifier. A comparison of models built from various combinations of the four basic fields showed a best performance when incorporating all these four fields, indicating that both profile information and tweet content are vital in inferring Twitter users' gender. Finally, M. Pennacchiotti and A.-M. Popescu[69] achieved their purpose of political orientation detection, ethnicity identification and business(exemplified with Starbucks) affinity detection by first applying a machine learning algorithm that classifies users based on profile, linguistic, behavioral and social features, and then a graph-based updating component that updates classification results based on the class label distribution of users' friends. Here, profile features are extracted from both information available on Twitter user's profile and other derived information such as age, gender and ethnicity; linguistic features are topical words, sentimental words, prototypical words and hashtags automatically extracted with a probabilistic model; behavioral features are a set of statistics indicating how the user interacts with others, such as tweeting frequency; and social features are statistics characterizing the user's social network, such as who he follows or whose tweets he retweets. Overall, M. Pennacchiotti and A.-M. Popescu's work was a successful combination of all the three approaches discussed in the above subsections.

4.2 User interest extraction

With the rapid growth of Twitter, users might soon find themselves overwhelmed by the huge amount of information every day, and thus would like to retrieve tweets that are indeed of their interest. This user need puts the accurate method to discover users' topics of interest at high priority. According to the different expressing methods of users' interest, we specifically divide the research into two categories: Term-Level Approach and Category-Level Approach, which we detail below.

4.2.1 Term-Level approaches

In the Term-Level Approach, user interest is described with words that are frequently used in users' tweets or to distinguish different user groups. Jilin Chen et al.[15] built a bag-of-words profile for each Twitter user from tweet streams to model his topics of interest. In their approach, the interestingness of each word was measured by the TF-IDF technique, based on the intuition that TF is a good indicator of how frequently users mention the word, and that IDF is a good indicator of the word's ability to distinguish one user from other users. As a user's own tweets, as well as his followees' tweets can both to some extent reflect the user's interest, Jilin Chen et al. built user profiles from each of the above sources, referred to as Self-Profile and Followee-Profile separately. According to their experiments, Self-Profile can indeed capture a user's interest as an information producer, while the Followee-Profile helps indicate his interest as an information seeker.

Though Jilin Chen et al.'s work has sort of achieved accuracy, there are still several natural limitations of messages in Twitter that researchers should take into account when modeling users' interest. For example, tweets have a length restriction of up to 140 characters, which is substantially different from documents in the traditional information retrieval area, and thus making techniques such as TF-IDF inappropriate to use. Consequently, researchers thereafter began to utilize Topic-Models(e.g. LDA, Latent Dirichlet Allocation) which are powerful tools to identify latent patterns in textual content. To discover topics in which Twitter users are interested by using LDA, the "unit" in Twitter that corresponds to document in traditional LDA should be firstly defined. L. Hong and B. D. Davison[45] compared the performance of three different defining metrics: 1) MSG – Each tweet was defined as a document, and LDA was trained to extract the topics of each tweet. Then, topics extracted from all tweets posted by the same user were aggregated to serve as the interest of that user; 2) USER – Tweets posted by the same user were aggregated into a document, and LDA was trained to extract the topics of that document, which were treated as the interest of that user; 3) TERM – Tweets containing a particular term were aggregated into a document, and LDA was trained to extract the topics of that document. Then, topics extracted from all terms contained in a user's tweet vocabulary were aggregated to serve as the interest of that user. L. Hong conducted an empirical study on the above three metrics, among which USER was revealed to be best performing. The result also assures the precision of J. Weng et al.'s[96] work, in which they developed an algorithm named TwitterRank to detect topic-sensitive influential twitters, using interest similarity between users as one of the measurements of influence. In their work, an Author-Topic Model similar to USER was utilized to extract users' topics of interest from tweets.

The works discussed above all leverage users' whole tweets collection to model user interest. However, users often use Twitter to initiate social chatting or tweet about their daily life, which indicates that a huge amount of tweets do not necessarily reflect user's interest, thus bringing great noise into the work of modeling user interest. Z. Xu et al.[99] addressed this issue by a Modified Author-Topic Model. They incorporated a latent variable into the original Author-Topic Model used by J. Weng et al. to indicate whether or not a tweet is related to its author's interest, and then conduct a tweet-level selection to distinguish tweets between user interest and social needs. They found that such interactions between users as retweet and reply are strong signals of common interest and thus using the existence of retweet, reply and url in tweet as an indicator of interest-relatedness of that tweet. Their experimental results on a manually built Twitter dataset verified that a better understanding of users' topics of interest can be reached by capturing the real motivation of tweets.

With the recently added Twitter capability for users to create lists of their friends, D. Kim et al.[53] tried to further discover user interest from the publicly available lists. They used χ^2 feature selection on the tweets of a particular list to find representative words that differentiated one list from other lists, which they defined as interest of the users in that list. Their user study confirmed that the words extracted from each list are good indicators of the interest of all the users in that list, even those who do not tweet about these words.

4.2.2 Category-Level approaches

The Term-Level Approaches referred to above are sometimes meaningless and may not be appropriate for clustering users or searching users by high-level topics. For instance, a user interested in NBA Games may never find LA Lakers' tweets if he only writes about Houston Rockets and its players specifically, without mentioning the term "NBA basketball club", which therefore won't appear in the abovementioned TF-IDF or LDA topic. Thus, a better way to model users' topics of interest is to use semantic representations such as Named Entities or even Categories, namely, the Category-Level Approach.

S. Piao and J. Whittle[70] constructed a realtime system to automatically identify users' topics of interest in the form of named entities and core terms by leveraging a series of NLP(Natural Language Processing) tools, such as POS tagger and Named Entity Recognizer. Though S. Piao and J. Whittle's work has achieved some sort of accuracy, the 140 characters restriction of tweets has rendered them extremely ungrammatical, noisy and full of abbreviations, which makes NLP not so appropriate a technique to extract named entities from tweets. Instead, M. Michelson and S. A. Macskassy[63] treated all capitalized, non-stopwords as candidate named entities, and leveraged Wikipedia as a knowledge base for disambiguation based on the context(words) around the discovered named entities. Once disambiguated, the entities were mapped to the categories contained in "folksonomy", a Wikipedia user-defined category tree, which were finally treated as users' topics of interest. Their experimental results showed that the usage of external knowledge bases such as Wikipedia could significantly empower entity disambiguation and category matching to generate reasonable topic profiles for Twitter users. Inspired by this result, R. Pochampally and V. Varma[71] also leveraged Wikipedia's category structure to conduct disambiguation and category matching. However, instead of only using a user's own tweets as the source for interest mining which was the idea of M. Michelson and S. A. Macskassy's work, R. Pochampally and V. Varma focused on tweets and lists metadata of user groups assumed to have common interest, and detected based on user context which was extracted from conversation patterns and Twitter list co-occurrence. They defined conversation patterns in Twitter as having three components: 1) proportion of user mentions (addressing or replying-to a user with '@'), 2) proportion of retweets(reposting another user's tweets with 'RT'), and 3) frequency of following(subscribing to a user's tweets). Their work achieved good performance in spite of the fact that tweets are ambiguous and noisy, and successfully extracted users' interests which were not evident from tweets alone.

4.3 Application

Personal information extracted from Twitter can be of great use in many applications such as demographic analyses of Twitter, personalized information services, location based services and user/tweet recommendations, to name a few. In this subsection, we briefly summarize some of the related works below.

`Demographic Analysis.` There have been several papers reporting on Twitter's demographic features in its entirety. Pear Analysis[1] estimated that as of August 2009, Twitter's population was composed of 55% female and 45% male. A consistent result was also obtained in J. D. Burger et al.'s initial manual analysis[10] using labels derived from blogs in users' profiles to determine gender, and B. Heil and M. J. Piskorski's work[42] who used name/gender correlations as a gender indicator. Pear Analysis[1] also found that 43% of Twitter's users are 18-34 years old, in contrast to users of Facebook and MySpace who are younger. This might partially be due to the fact that unlike Facebook and MySpace in which users' requests to subscribe to others need to be authorized, and thus only reciprocal relationships are allowed, Twitter's following network requires no reciprocation, which leads to poorer security and less frequent adoption among younger people.[19] Other research mainly focused on Twitter's geographical properties, among which A. Java et al.'s work[50] was the most comprehensive and influential. They found that: Twitter is most widely adopted in US, Europe and Asia (mainly Japan); Europeans and Asians have a higher tendency to connect to others who speak the same language; and there are not so many across-continent friendships in Twitter compared to those intra-continent.

`Content Customization.` With the increasing popularity of Twitter, huge amounts of information are poured into it every day, and users might soon find themselves overwhelmed by the flood of tweets. This phenomenon was confirmed by M. Bernstein et al.[5] who conducted an informal survey among 78 participants aged 14-47, and found that the average number of daily tweets for these users is about 786, which is obviously beyond a user's consumption per day, and that users always prefer tweets which are highly relevant to their interest. So they developed an alternative Twitter client called "Eddi" which clusters a user's tweets into various topics and only displays those according with his preference or interest.

5 Social Information Extraction

Social Information can be defined as information that identifies users' relationships or interactions with others, including topological structure, community distribution and even detailed social relationships such as co-workers, family members and so on. In this section, we first outline Twitter's topological characteristics such as degree distribution, reciprocity, homophily and hybrid network feature, according to a series of existing topological analysis works, and then address the problem of link prediction, or friend recommendation, in the context of the Twitter community.

5.1 Topological analysis

As a combination of social networking and an information service, Twitter was born with many topological features that distinguish itself from traditional networks. On the one hand, unlike other social networking sites such as Facebook and MySpace which are often described as undirected graphs, the Twitter network can be modeled as a directed graph $G = \{V, E\}$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of users, and $E \subseteq V \times V$ is the set of directed edges between users. When user v_i choose to follow user v_j, there is a directed edge from v_i to v_j, donated as $e_{ij} = (v_i, v_j) \in E$; On the other hand, unlike the pure information service on the Web, parts of the Twitter network do parallel with the offline friendship between users. Below, we will detail these distinct network features of Twitter.

`Hybrid Network Feature.` Users on Twitter subscribe to other users' tweets by following actions, which constitute an explicit network structure. However, because of the scarcity of user attention, this huge network cannot reveal actual interactions among people. There is an implicit network consisting of communications between users and their actual friends. By defining a user's actual friend as a person with whom the user has initiated at least two directed conversations(i.e. '@' followed by user identifier), B. A. Huberman et al.[47] found out that the implicit network of actual friends is much sparser than the explicit network made up of followees and followers. M. J. Welch and U. Schonfeld[95] considered communications between users as retweet actions and concluded according to their experimental results that these retweet edges are significantly better for preserving topical relevance than following edges. Finally, J. Hopcroft et al.[46] in their work to predict reciprocal relationships in Twitter, found that the implicit network of retweet or reply edges are more related to the formation of reciprocal relationships.

`Degree Distribution.` A. Java et al.[49] found that the majority of their crawled Twitter data exists with a high degree correlation, that is, users with many followees achieve a higher tendency to have many followers. And by calculating the cumulative degree distributions of the Twitter network, they also noted that the in-degree and out-degree satisfy power-law distribution with an exponential of -2.4. Similar results were discovered in M. J. Welch and U. Schonfeld's work[95], in which they noted that the explicit following network indeed has a power-law distribution, while the implicit network of retweet edges shows power-law distribution in in-links but a non-power-law distribution in out-links. B. Krishnamurthy et al.[54], on the other hand, discovered a non-power-law degree distribution in their crawled Twitter following network, which was further confirmed by A. Java et al.[49] who conducted a topological study on the entire Twittersphere. B. Krishnamurthy et al.[54] also divided Twitter users into three categories based on their in-degree/out-degree ratio: broadcaster with a high in-degree but a low out-degree; miscreant(spammer or stalker) with a low in-degree but a high out-degree; and acquaintance with high degree correlation and high reciprocity. Finally, A. Ronel and M. Teutle[91], who studied Twitter network dynamics, found out that users with more than 600 followers increase their in-degree rapidly, indicating a "rich gets richer" phenomenon, while out-degree didn't change as much as in-degree.

`Homophily.` Homophily is "a tendency that contact between similar users occurs at a higher rate than among dissimilar users"[55]. A great deal of research has revealed homophily in Twitter, including location homophily, status homophily and link homophily. A. Java et al.[49]'s work was the first to study relationships between structural properties in the Twitter network and geographic properties in the physical world. They found that there are not so much across-continent friendships as those intra-continent, indicating that the probability of friendship between two users is inversely proportional to their geographic distance. J. Hopcroft et al.'s[46] quantitative experiment also showed that users from the same time zone have a 50 times higher tendency to achieve reciprocal relationships than those with a distance of three time zones away. This location homophily feature in Twitter was further confirmed by H. Kwak et al.[55] and S. Yardi and D. Boyd[100]. S. Yardi and D. Boyd even found that the geographically local networks are denser and more connected than the nonlocal ones, and that central users on a specific topic are also centrally located in the physical world. H. Kwak et al.[55] and J. Hopcroft et al.[46]'s work revealed the phenomenon of status homophily, that is the tendency of celebrities(i.e., users with abundant followers) to follow each other is much stronger than that of ordinary users, which is even stronger than the tendency of relationships between celebrities and ordinary users. J. Hopcroft et al.[46] also discovered that users with more common reciprocal friends have a much higher likelihood to follow each other, which we refer to as link homophily here.

`Reciprocity.` A. Java et al.[49] and A. Ronel and M. Teutle[91] found a high reciprocity in the Twitter network(around 50% of the users have a two-way relationship in the latter's collected dataset) especially in Asian and European communities, indicating close mutual acquaintances among users in these areas. However, H. Kwak et al.'s[55] topological analysis on the entire Twittersphere demonstrated that only 22.1% of users own two-way relationships, reflecting a low reciprocity. As for the factors that affect the extent of reciprocity, B. Krishnamurthy et al.[54] found that users tweeting frequently tend to have more reciprocal relationships; and M. S. Smith and C. G. Carrier[87] also found that Twitter users who request to follow others with similar interest tremendously increase the number of reciprocate responses.

5.2 Link prediction

Link prediction is the problem of recommending potential friends for Twitter users beyond the current network status. According to the information sources used, we specifically divide existing works on link prediction into three categories: Content-Based Approach, Network-Based Approach and Hybrid Approach, which we detail below.

5.2.1 Content-Based approaches

By analyzing data collected between early February and the end of March 2010, D. Yin et al.[103] found that 90% of new links in Twitter are formed in the two-degree network, i.e., friends of friends, including both one-way and two-way(reciprocal) relationships. This result is often referred to as two-degree separation, and has been leveraged in most works related to link prediction. For example, T. Sakaguchi et al.[82] considered users within two-degree of separation to the original user as potential friends, from which they randomly chose 20 users as candidate set. They also applied linguistic analysis to the original user's tweets, extracted TF-IDF of each noun as user word vector, and used 15 words with a higher TF to search for people who have tweeted with these words recently, from which the top 20 users were chosen as another candidate set. The most prominent idea of T. Sakaguchi et al.'s work was the utilization of Concept Fuzzy Sets, which were based on "the use theory of meaning proposed by Witgenstein to express the meaning of concepts"[82]. They manually transformed each of the 450,000 articles in Japanese Wikipedia into a prototype concept represented as a word vector of TF-IDF values of a maximum of 40 words, and then expanded the original and candidate users' word vectors by the 50 most similar prototype concepts, and used a cosine measure to calculate the degree of similarity between users, thus ranking them for recommendation. Their experimental results showed a satisfactory performance, which confirmed the potential of leveraging tweets content as a source for friends recommendation.

5.2.2 Network-Based approaches

Network-Based Approaches predict potential links purely based on the current network status and structural evolution over time. For instance, S. A. Golder et al.[29] leveraged the phenomenon of link homophily, i.e., users with more common friends(be it one-way or two-way relationship) have a much higher likelihood to follow each other, in their friends recommendation strategy. They analyzed several principles for link prediction, such as shared interests(considering the number of common followees the original and recommended users follow), shared audience(considering the number of common followers the original and recommended users have), transitivity(considering potential friends as users followed by the users the original user is already following, i.e., followees of followees), and mutuality (or reciprocity, i.e. two-way relationship). Besides the abovementioned structural information, A. Golder et al.'s work one year later[30] also took into consideration profile information such as followee number, follower number, tweet number, account age of the original and recommended users, as a basis for predictions. They conducted a web-based user study in which they provided subjects with users within a two-degree network as candidate friends and asked them to rate their desire in forming friendships with those users. They then used a hierarchical regressing model to estimate the effects of both profile and structural characteristics on a subject's preference. The conclusions drawn from their experimental results were as follows: 1) users tweeting frequently are less desired to add new friends; 2) users with many followers already appear more likely to obtain new followers; 3) shared followees and shared followers don't have a statistically significant effect on users' tendencies to follow each other; 4) transitivity and mutuality are altogether important conditions in increasing users' desires to form friendships.

D. M. Romero and J. Kleinberg[80] employed the idea of directed closure process, a variation of triadic closure in traditional social networks, in their link prediction strategy. Triadic closure in such social networks as Facebook and MySpace is the assumption that two people who already have a friend in common are more likely to form a friendship, which is one of the fundamental processes of link formation. Analogously, users in Twitter also expect an increased likelihood to follow people their followees already follow. This idea was referred to by D. M. Romero and J. Kleinberg as "directed closure process" in their work. To verify this assumption in Twitter, they applied the preferential attachment method to a collected random sample of celebrities on Twitter and determined the subset of directed edges to a celebrity that exhibit closure to calculate his closure ratio. Their experimental results indicated that the directed closure process is indeed prevalent in the Twitter network, and that the more important factor to determine a user's closure ratio is the total number of followers of those following the user, rather than the number of followers the user himself has, or to be more precise, a user's closure ratio is more closely correlated with the sum of in-degrees of the followers from his own community than that of all his followers. D. Yin et al.[102] proposed a novel structure-based approach to link prediction. To calculate the likelihood of forming a directed link from user v_u to user v_c, for each intermediate user v_i, they combined the link structure between v_u and v_i, standing for the probability of v_u trusting the recommendation of v_i, and the link structure between v_i and v_c, standing for the probability of v_i recommending v_c. Unlike D. M. Romero and J. Kleinberg who only considered the followee-of-followee relationship[80], D. Yin et al. combined both the one-way relationship in each direction and the two-way relationship in a user's two-degree network. Their experimental comparison with many popular link prediction methods such as Common neighbors (simply considers the number of common friends), Jaccard coefficient(divides the number of common friends by the total number of friends), Adamic/Adar(weights rarer features more heavily), Preferential attachment(multiplies v_c's in-degree and v_u's out-degree), L. Katz's method[52], PropFlow[57], and the matrix factorization popular in recommender systems showed that their proposed model outperformed state-of-the-art methods on the link prediction task on Twitter.

Unlike the works discussed above, J. Hopcroft et al.[46] focused specifically on the prediction of reciprocal relationship. They proposed a learning framework and formulated the task into a graphical model, the Triad Factor Graph(TriFG) model. They showed that homophily(including location, link and status homophily introduced above) and structural balance("for every group of three users called triad, either all three of these users are friends or only one pair of them are friends"[46]) are prevalent in the reciprocal Twitter network, and that by incorporating such social theories as homophily and structural balance into the proposed TriFG model, the performance of reciprocal link prediction can be significantly improved(TriFG can accurately infer 90% of reciprocal friends in Twitter).

There have also been a great deal of third-party applications which incorporate Network-Based Approaches to enhance Twitter's ability to personalize friend recommendations. For example, MrTweet.com[5] (closed now) suggests potential followees

[5] http://mrtweet.com/

for users based on common friends; Twubble.com[6] (unavailable now due to lack of OAuth support) considers users within two-degree separation as candidate friends and ranks them based on the number of reciprocal relationships.

5.2.3 Hybrid approaches

Based on the social-information feature of Twitter, it can be concluded that "participants in Twitter create links for multiple reasons–to be social(i.e., to connect online to existing offline social contacts) or to link to information sources"[103]. So the link prediction task should accordingly consider two viewpoints–to recommend friends to users as in traditional social networks or to recommend an information source to an information consumer as in information networks, thus making the Hybrid Approach, which combines both network structure and tweet content, a better choice for link prediction in Twitter.

To verify this statement, K. Puniyani et al.[72] treated all the tweets posted by a single user as a document, applied supervised LDA to the document to learn the latent topics characterizing that user's interest, and then used a regression-based predictor to calculate the strength of connection between users for recommendation ranking. They evaluated this topic-based approach for link prediction on both the explicit following network and the implicit communicating network made up of user mentions ('@'), and compared its performance against the Common neighbors method which only considers network structure. The experimental results showed that the link-based approach is especially effective on the explicit following network, while the topic-based approach is more competitive on the implicit communicating network. J. Hannon et al.'s work[39] took this attempt a step further by considering both the user's own tweets and the tweets from his followees and followers as content-based filtering, and all the user's followees' and followers' IDs as collaborative filtering. They utilized Lucene's TF-IDF weighting metric to calculate users' content vocabulary features and structure vocabulary features. According to their off-line evaluation and a live-user trial(Twittomender), structure-based strategies perform better in recommendation precision, while content-based strategies perform better in ranking effectiveness.

Based on these observations, A. Sadilek et al.[81] incorporated content information such as text-similarity quantifying the amount of overlaps in users' tweeting vocabulary, co-location measures extracted from users' tweets which capture how often users tend to stay close to each other, as well as structure information such as common friends ratio, into their recommendation model. Their empirical study showed a good predicting performance even with no previously observed links, due to the utilization of content features.

6 Travel Information Extraction

Travel Information can be defined as information about users' current location, travel history, or whether or not a specific user would be on vacation away from his house and when the vacation would be. In this section, we first summarize existing research in the area of travel information extraction from Twitter such as current location detection and future vacation detection. Then we present a brief introduction

[6] `http://crazybob.org/twubble/`

of applications where the extracted travel information can be utilized.

6.1 Current location detection

With the recent development of Location Based Services(LBSs), mining users' current location promises new crisis management technologies and personalized services, including local event detection and regional targeted advertising. The simplest way is to leverage information contained in users' profiles to determine their current locations. For example, "Location" values of profiles contained in the metadata of tweets can be utilized to locate Twitter users. This information is always expressed as specific place names, but when GPS-enabled mobile devices are used, more precise locations can be identified with GPS coordinates. M. Cheong and V. Lee[17,18], T. Sakaki et al.[83], M. Guy et al.[36] and A. Java et al.[50] all used place names in the "Location" field as users' current location, while the former four also resorted to GPS coordinates to generate more accurate results. Besides the "Location" values, there still are other profile information that can be utilized to infer users' current location. For instance, B. Krishnamurthy et al.[54] used UTC(Coordinated Universal Time) offset in the timezone field of tweets and URL domain names(e.g. .com for US, .jp for Japan, .de for Germany and .uk for UK) in profile "URL" field as their bases for location mining.

The abovementioned approaches all focused purely on information contained in users' or tweets' profiles, which intuitively could not achieve good enough performance. B. Hecht et al.[41] on the other hand, paid attention to the content of tweets and developed a Multinomial Naïve Bayes(MNB) model to train a classifier to determine a user's current location. The input to this model was expressed as a term vector with each dimension representing a term in a user's tweeting vocabulary and the value of the dimension representing the TF of that term. Their experimental results indicated high accuracy in determining a user's country-level and state-level location. B. D. Longueville et al.[21] took this strategy a step further by using the "Location" field and GPS coordinates contained in a user's profile, as well as place names, URL domains and hashtags of places contained in tweets, as his current location indicators.

Previous works all rely on the precondition that either user profiles or tweets are public and available for location prediction, which is not always the case. To address this, A. Sadilek et al.[81], inspired by the intuition that friends often participate in activities together in a specific place, implemented a system called "Flap", which infers users' current locations based on known GPS positions of their friends. They formulated the problem of location prediction as a dynamic Bayesian network(DBN) with one hidden node representing the location of the target user and a number of observed nodes representing the locations of the target user's friends in each time slice(20 minutes). They also incorporated into the model an observed node representing the time of day and another observed node determining whether a given day is a work day or a free day(e.g., weekend or a national holiday). They used both supervised and unsupervised learning to train DBN which was then used to infer the most likely sequence of locations one visited(i.e., the value of the hidden node) over a specific time period, given a sequence of locations visited by his friends, along with the corresponding time and day type. Their experimental results demonstrated that by leveraging social ties, the proposed DBN model can infer users' locations with high accuracy and fine granularity in both space and time even when users keep their tweets and GPS data private.

6.2 Future vacation detection

H. Mao et al.'s work[60] is the only one we find that amply addressed the problem of future vacation detection. They firstly acquired real-time Twitter data via the Twitter Streaming API and applied keyword matching to these data to filter out all potentially vacation-related tweets. Examples of topical keywords included "vacation", "holiday", "travel", "trip", "leave for" and "fly to". They then leveraged Naïve Bayes and SVM to classify each tweet as sensitive(i.e., reveal users' travel plans) or non-sensitive. The representative words they used as classification features were derived from an analysis of 1,000 randomly sampled and manually annotated tweets. They found that among the 108 vacation sensitive tweets, 90.7%, 55.5%, and 44.4% have an occurrence of location, time and person, respectively, so they used After NER and Alchemy NER to automatically detect location, person and time(LPT) mentions in candidate tweets. In addition to those LPT features, they also found that "some places representative of vacation(e.g. beach, coast, hotel) as well as air transportation(e.g. airport, flight), and some words implying preparation for vacation(e.g. leave, pack, book, plan) are good indicators of vacation sensitive tweets, while other words, including negative words (e.g. not, no, didn't), virtual words(e.g. should, wish, need, if) as well as past tense verbs(e.g. went, got) implying that the travel is already past or is not real, are good indicators of non-sensitive tweets"[60]. Therefore, they also chose these representative words and phrases as classification features. Their experimental results demonstrated a 76% precision in future vacation detection.

6.3 Application

Travel information extracted from Twitter can be of great use in many applications such as location based services, crisis management and burglary defense, to name a few. In this subsection, we briefly summarize some of the related works below.

`Crisis Management.` Twitter is widely used as a realtime information dissemination platform, which makes itself a potential tool for monitoring and managing crisis and convergence events, if accurate spatiotemporal information related to the events can be derived. B. D. Longueville et al.[21] tried this usage by mining tweets to track forest fires in Marseille, France and found that "the timeline of tweets did accurately match the real-world spread of the fire, except for a lag time at the beginning of the fire". T. Sakaki et al.[83] and M. Guy et al.[36] took this application a step further by treating Twitter users as social sensors to realize early detection and warning of potential emergent situations. T. Sakaki et al.[83] constructed a prototype earthquake reporting system which could detect earthquakes by picking out earthquake-related messages from Japanese language tweets stream and then applying Kalman and Particle filters to spatiotemporal information in tweets to predict the trajectories of the earthquakes. Similar to this work, M. Guy et al.[36], researchers in the US Geological Society, also developed a global earthquake detecting system called "Twitter Earthquake Detector"(TED). Though there are still several unsolved technical issues in the system now, its superiority over traditional detectors has been confirmed by an almost four times faster detection of an earthquake in Indonesia.

`Burglary Defense.` While privacy issues have obtained rising awareness on social networks such as Twitter and some users do change their privacy settings to "private", there still are a larger amount of users who expose their profiles and tweets to the public or even unwittingly reveal their vacation plans in tweets, which makes them vulnerable to theft. Actually, Twitter users have already been burglarized in this way[64]. Fortunately, some geolocating applications have been developed to raise awareness about the lack of location privacy in Twitter. For example, ICanStalkU.com[7] leverages photo information shared in Twitter to extract users' current locations even without them realizing it. PleaseRobMe.com[8] scans users' public tweets streams for location-related messages, and then uses Foursquare's GPS-enabled mobile devices to extracted their geographic check-ins which, if inconsistent with their registered home address, might lead to burglaries. Additionally, the classification strategy utilized by H. Mao et al[60] can also be applied to future "guardian angel systems", which would monitor users' tweets and alert them of potential leak of vacation plans, thus strengthening burglary defense.

7 Discussion

The prior sections have highlighted the current state of the art in information extraction from Twitter. In Table 1, we summarize the findings of prior sections, and also list noteworthy papers discussed.

Table 1 Summary of research covered

	Profile-Based	Content-Based	Network-Based	Hybrid
Personal	[42][18][17][69]	[48][60][16][74][15][45][96] [99][53][70][63][71]	[62][28]	[69][66][10]
Social		[82]	[46][29][30][80][102]	[72][39][81]
Travel	[18][17][50][54][83][36]	[60][14]		[81][21]

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 this field to get started quickly.

1) how to extract information from microblogs. As we mentioned before, 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. We note that all these three types of sources are actually related to each other. In other words, we can use any of them to estimate the other two. For example, we can infer users' interests from their tweets, or recommend new followees based on users' interests, etc. For now, the authors of this paper are exploring the approaches to combine users' mobilities with the social network between them, namely the so-called *social media mobility* problem. Or more specifically, we are attempting to conduct friend recommendation based on users' co-location patterns. We consider it to be a trending topic in the near future, with the rising prevalence of location-based services.

[7] `http://icanstalkyou.com/`

[8] `http://pleaserobme.com/`

2) 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 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.

3) how to guarantee privacy. As more and more sensitive information is being extracted from Twitter, users, even those who use protected accounts(which would limit access to friends only), will soon find themselves facing serious problems of privacy leaks. As a consequence, how to guarantee users' privacy has become an relatively urgent issue recently. This also promises future research opportunities to build guardian services which can automatically detect privacy leaks in Twitter and then inform users to "think twice before tweeting".

8 Conclusion

This paper has reviewed the recent state of the art in the literature of information extraction from microblogging services(exemplified with Twitter). We specifically focus on three types of information – personal, social and travel information, discuss prevalent approaches used in each area, as well as their applications in daily life, and then propose some suggestions for future work. In our opinion, this paper can serve as a guidance to researchers interested in this field.

References

[1] Pear analytics: Twitter study august 2009. http://www.pearanalytics.com/blog/wpcontent/uploads/2010/05/twitter-study-august-2009.pdf. [Technical Report] August 2009.

[2] Agichtein E, Gravano L, Pavel J, Sokolova V, Voskoboynik A. Snowball: a prototype system for extracting relations from large text collections. Proc. of the 2001 ACM SIGMOD International Conference on Management of Data. SIGMOD '01. ACM. New York, NY, USA. 2001.

[3] Banko M, Cafarella M, Soderland S, Broadhead M, Etzioni O. Open information extraction from the web. Proc. of IJCAI. 2007.

[4] Benson E, Haghighi A, Barzilay R. Event discovery in social media feeds. Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. HLT '11. Association for Computational Linguistics. Stroudsburg, PA, USA. 2011. 389–398.

[5] Bernstein M, Hong L, Kairam S, Chi H, Suh B. A torrent of tweets: Managing information overload in online social streams. Conference on Human Factors in Computing Systems. CHI '10. 2010.

[6] Bikel DM, Schwartz R, Weischedel RM. An algorithm that learns whatś in a name. Mach. Learn, February 1999, 34: 211–231.

[7] Bollegala D, Matsuo Y, Ishizuka M. Relational duality: unsupervised extraction of semantic relations between entities on the web. Proc. of the 19th International Conference on World Wide Web. ACM. 2010. 151–160.

[8] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Proc. of the seventh international conference on World Wide Web 7. WWW7, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V. 1998. 107–117.

[9] Bunescu RC, Mooney RJ. A shortest path dependency kernel for relation extraction. HLT '05: Proc. of the conference on Human Language Technology and Empirical Methods in Natural

Language Processing. Association for Computational Linguistics. Morristown, NJ, USA. 2005. 724–731.

[10] Burger J, Henderson J, Kim G, Zarrella G. Discriminating gender on twitter. Applying Systems Engineering and Advanced Technology to Critical National Promblems. 2011.

[11] Cafarella MJ, Halevy A, Khoussainova N. Data integration for the relational web. Proc. of the VLDB Endow. August 2009(2): 1090–1101.

[12] Cafarella MJ, Halevy A, Wang DZ, Wu E, Zhang Y. Webtables: exploring the power of tables on the web. Proc. of the VLDB Endow. August 2008(1): 538–549.

[13] Cafarella MJ, Madhavan J, Halevy A. Web-scale extraction of structured data. SIGMOD Rec, 2008, 37(4):55–61.

[14] Carlson A, Betteridge J, Kisiel B, Settles B, Jr. ERH, Mitchell TM. Toward an architecture for never-ending language learning. Proc. of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010). 2010.

[15] Chen J, Nairn R, Nelson L, Bernstein M, Chi E. Short and tweet: experiments on recommending content from information streams. Proc. of the 28th International Conference on Human Factors in Computing Systems. CHI '10. ACM. New York, NY, USA. 2010. 1185–1194.

[16] Cheng Z, Caverlee J, Lee K. You are where you tweet: a content-based approach to geo-locating twitter users. Proc. of the 19th ACM International Conference on Information and Knowledge Management. CIKM '10. ACM. New York, NY, USA. 2010. 759–768.

[17] Cheong M, Lee V. Integrating web-based intelligence retrieval and decisionmaking from the twitter trends knowledge base. Proc. of the 2nd ACM Workshop on Social Web Search and Mining. SWSM '09. ACM. New York, NY, USA. 2009. 1–8.

[18] Cheong M, Lee V. Twitmographics: Learning the emergent properties of the twitter community. In: Memon N, Alhajj R, eds. From Sociology to Computing in Social Networks. Springer Vienna, 2010. 323–342. doi: 10.1007/978-3-7091-0294-7_17.

[19] Cheong M, Ray S. A literature review of recent microblogging developments. Techinical Report. 2011.

[20] Cunningham H, Maynard D, Bontcheva K, Tablan V. Gate: A framework and graphical development environment for robust nlp tools and applications. Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics. 2002.

[21] De Longueville B, Smith RS, Luraschi G. "omg, from here, i can see the ames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. Proc. of the 2009 International Workshop on Location Based Social Networks. LBSN '09. ACM. New York, NY, USA. 2009. 73–80.

[22] Efron M. The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents. Proc. of the thirteenth ACM international conference on Information and knowledge management. CIKM '04. ACM. New York, NY, USA. 2004. 390–398.

[23] Elmeleegy H, Madhavan J, Halevy A. Harvesting relational tables from lists on the web. Proc. of the VLDB Endow. August 2009, 2: 1078–1089.

[24] Etzioni O, Cafarella M, Downey D, Kok S, Popescu AM, Shaked T, Soderland S,Weld DS, Yates A. Web-scale information extraction in knowitall. Proc. of the 13th International Conference on World Wide Web. WWW '04. ACM. New York, NY, USA. 2004. 100–110.

[25] Etzioni O, Cafarella M, Downey D, Popescu AM, Shaked T, Soderland S, Weld DS, Yates A. Unsupervised named-entity extraction from the web: an experimental study. Artificial Intelligence, June 2005, 165: 91–134.

[26] Finin T, Murnane W, Karandikar A, Keller N, Martineau J, Dredze M. Annotating named entities in twitter data with crowdsourcing. Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics. CSLDAMT '10. Stroudsburg, PA, USA. 2010. 80–88.

[27] Gimpel K, Schneider N, O'Connor B, Das D, Mills D, Eisenstein J, Heilman M, Yogatama D, Flanigan J, Smith NA. Part-of-speech tagging for twitter: annotation, features, and experiments. Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. HLT '11. Association for Computa-

tional Linguistics. Stroudsburg, PA, USA. 2011. 42–47.

[28] Golbeck J, Hansen D. Computing political preference among twitter followers. Proc. of the 2011 annual conference on Human factors in computing systems. CHI '11. ACM. New York, NY, USA. 2011. 1105–1108.

[29] Golder S, Marwick A, Yardi S. A structural approach to contact recommendations in online social networks. Proc. of 2nd Workshop on Search in Social Media. SSM '09. 2009.

[30] Golder S. Yardi S. Structural predictors of tie formation in twitter: Transitivity and mutuality. Social Computing (SocialCom), 2010 IEEE Second International Conference on. Aug. 2010. 88–95.

[31] Gouws S, Metzler D, Cai C, Hovy E. Contextual bearing on linguistic variation in social media. Proc. of the Workshop on Languages in Social Media. LSM '11. Association for Computational Linguistics. Stroudsburg, PA, USA. 2011. 20–29.

[32] Grishman R, Sundheim B. Message understanding conference-6: a brief history. Proc. of the 16th conference on Computational linguistics-Volume 1. COL-ING '96. Association for Computational Linguistics. Stroudsburg, PA, USA. 1996. 466–471.

[33] Gulhane P, Rastogi R, Sengamedu SH, Tengli A. Exploiting content redundancy for web information extraction. Proc. of the 19th international conference on World wide web. WWW '10. ACM. New York, NY, USA. 2010. 1105–1106.

[34] Guo J, Xu G, Cheng X, Li H. Named entity recognition in query. SIGIR '09: Proc. of the 32nd international ACM SIGIR conference on Research and Development in Information Retrieval. ACM. New York, NY, USA. 2009. 267–274.

[35] Gupta R, Sarawagi S. Answering table augmentation queries from unstructured lists on the web. Proc. of the VLDB Endow. August 2009, 2:289–300.

[36] Guy M, Earle P, Ostrum C, Gruchalla K, Horvath S. Integration and dissem- ination of citizen reported and seismically derived earthquake information via social network technologies. In: Cohen P, Adams N, Berthold M, eds. Advances in Intelligent Data Analysis IX. LNCS 6065. Springer Berlin/Heidelberg, 2010. 42–53. doi: 10.1007/978-3-642-13062-5 6.

[37] Han B, Baldwin T. Lexical normalisation of short text messages: makn sens a #twitter. Procs. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. HLT '11. Association for Computational Linguistics. Stroudsburg, PA, USA. 2011. 368–378.

[38] Han H, Giles CL, Manavoglu E, Zha H, Zhang Z, Fox EA. Automatic document metadata extraction using support vector machines. Proc. of the 3rd ACM/IEEE-CS Joint Conference on Digital libraries. 2003. 37–48.

[39] Hannon J, Bennett M, Smyth B. Recommending twitter users to follow using content and collaborative filtering approaches. Proc. of the Fourth ACM Conference on Recommender Systems. RecSys '10. ACM. New York, NY, USA. 2010. 199–206.

[40] Hearst MA. Automatic acquisition of hyponyms from large text corpora. Proc. of the 14th Conference on Computational Linguistics. Association for Computational Linguistics. Morristown, NJ, USA. 1992. 539–545.

[41] Hecht B, Hong L, Suh B, Chi EH. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. Proc. of the 2011 Annual Conference on Human Factors in Computing Systems. CHI '11. ACM. New York, NY, USA. 2011. 237–246.

[42] Heil B, Piskorski M. New twitter research: Men follow men and nobody tweets. http://blogs.hbr.org/cs/2009/06/new twitter research men follo.html.

[43] Hindle D. Noun classification from predicate-argument structures. Proc. of the 28th annual meeting on Association for Computational Linguistics. ACL '90. Association for Computational Linguistics. Stroudsburg, PA, USA. 1990. 268–275.

[44] Hobbs JR, Bear J, Israel D, Tyson M. Fastus: A finite-state processor for information extraction from real-world text. IJCAI. 1993. 1172–1178.

[45] Hong L, Davison BD. Empirical study of topic modeling in twitter. Proc. of the First Workshop on Social Media Analytics. SOMA '10. ACM. New York, NY, USA. 2010. 80–88.

[46] Hopcroft J, Lou T, Tang J. Who will follow you back?: reciprocal relationship prediction. Proc. of the 20th ACM International Conference on Information and Knowledge Management. CIKM

'11. ACM. New York, NY, USA. 2011. 1137–1146.

[47] Huberman B, Romero D, Wu F. Social networks that matter: Twitter under the microscope. First Monday. 2009, 14(1): 8.

[48] Humphreys L, Gill P, Krishnamurthy B. Privacy on twitter: How much is too much? privacy issues on twitter. The Annual Meeting of the International Communication Association. 2010.

[49] Java A, Song X, Finin T, Tseng B. Why we twitter: understanding microblogging usage and communities. Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. WebKDD/SNA-KDD '07. ACM. New York, NY, USA. 2007. 56–65.

[50] Java A, Song X, Finin T, Tseng B. Why we twitter: An analysis of a microblog- ging community. In: Zhang H, Spiliopoulou M, Mobasher B, Giles C, McCallum A, Nasraoui O, Srivastava J, Yen J, eds. Advances in Web Mining and Web Usage Analysis. LNCS 5439. Springer Berlin / Heidelberg. 2009. 118–138. doi: 10.1007/978-3-642-00528-2 7.

[51] Kasneci G, Ramanath M, Suchanek F, Weikum G. The yago-naga approach to knowledge discovery. SIGMOD Rec, 2008, 37(4):41–47.

[52] Katz L. A new status index derived from sociometric analysis. Psychometrika, 1953, 18: 39–43. doi: 10.1007/BF02289026.

[53] Kim D, Jo Y, Moon IC, Oh A. Analysis of twitter lists as a potential source for discovering latent characteristics of users. Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems. (CHI 2010). 2010.

[54] Krishnamurthy B, Gill P, Arlitt M. A few chirps about twitter. Proc. of the first workshop on Online social networks. WOSN '08. ACM. New York, NY, USA. 2008. 19–24.

[55] Kwak H, Lee C, Park H, Moon S. What is twitter, a social network or a news media? Proc. of the 19th international conference on World wide web. WWW '10. ACM. New York, NY, USA. 2010. 591–600.

[56] Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc. of the 18th International Conference on Machine Learning. 2001. 282–289.

[57] Lichtenwalter RN, Lussier JT, Chawla NV. New perspectives and methods in link prediction. Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '10. ACM. New York, NY, USA. 2010. 243–252.

[58] Limaye G, Sarawagi S, Chakrabarti S. Annotating and searching web tables using entities, types and relationships. Proc. of VLDB Endowment, September 2010, 3: 1338–1347.

[59] Liu X, Zhang S, Wei F, Zhou M. Recognizing named entities in tweets. Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. HLT '11. Association for Computational Linguistics. Stroudsburg, PA, USA. 2011. 359–367.

[60] Mao H, Shuai X, Kapadia A. Loose tweets: an analysis of privacy leaks on twitter. Proc. of the 10th Annual ACM Workshop on Privacy in the Electronic Society. WPES '11. ACM. New York, NY, USA. 2011. 1–12.

[61] McCallum A, Freitag D, Pereira FCN. Maximum entropy markov models for information extraction and segmentation. Proc. of the Seventeenth International Conference on Machine Learning. ICML '00. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA. 2000. 591–598.

[62] Metaxas P, Mustafaraj E. From obscurity to prominence in minutes: Political speech and real-time search. WebSci10: Extending the Frontiers of Society On-Line. `http://bit.ly/h3Mfld`. 2010.

[63] Michelson M, Macskassy SA. Discovering users' topics of interest on twitter: a first look. Proc. of the fourth workshop on Analytics for noisy unstructured text data. AND '10. ACM. New York, NY, USA. 2010. 73–80.

[64] Mills E. Twitter user says vacation tweets led to burglary. June 2008. http://news.cnet.com/8301-1009 3-10260183-83.html.

[65] Moschitti A. Efficient convolution kernels for dependency and constituent syntactic trees. Proc. of the ECML. 2006. 318–329.

[66] Mustafaraj E, Metaxas P. What edited retweets reveal about online political discourse. Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence. AAAI '11. 2011.

[67] Nguyen TVT, Moschitti A, Riccardi G. Convolution kernels on constituent, dependency and sequential structures for relation extraction. EMNLP '09: Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. Morristown, NJ, USA. 2009. 1378–1387.

[68] Peng F, McCallum A. Information extraction from research papers using conditional random fields. Information Processing and Management, July 2006, 42: 963–979.

[69] Pennacchiotti M, Popescu AM. Democrats, republicans and starbucks afficionados: user classification in twitter. In Proc. of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '11. ACM. New York, NY, USA. 2011. 430–438.

[70] Piao S, Whittle J. A feasibility study on extracting twitter users' interests using nlp tools for serendipitous connections. Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom). Oct. 2011. 910–915.

[71] Pochampally R, Varma V. User context as a source of topic retrieval in twitter. SIGIR 2011 Workshop on Enriching Information Retrieval. ENIR '11. 2011.

[72] Puniyani K, Eisenstein J, Cohen S, Xing EP. Social links from latent topics in microblogs. Proc. of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media. WSA '10. Association for Computational Linguistics. Stroudsburg, PA, USA. 2010. 19–20.

[73] Ramage D, Hall D, Nallapati R, Manning CD. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. EMNLP '09. Association for Computational Linguistics. Stroudsburg, PA, USA. 2009. 248–256.

[74] Rao D, Yarowsky D, Shreevats A, Gupta M. Classifying latent user attributes in twitter. Proc. of the 2nd International Workshop on Search and Mining User-Generated Contents. SMUC '10. ACM. New York, NY, USA. 2010. 37–44.

[75] Reichartz F, Korte H, Paass G. Composite kernels for relation extraction. ACL-IJCNLP '09: Proc. of the ACL-IJCNLP 2009 Conference Short Papers. Association for Computational Linguistics. Morristown, NJ, USA. 2009. 365–368.

[76] Reichartz F, Korte H, Paass G. Semantic relation extraction with kernels over typed dependency trees. KDD '10: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. New York, NY, USA. 2010. 773–782.

[77] Reiss F, Raghavan S, Krishnamurthy R, Zhu H, Vaithyanathan S. An algebraic approach to rule-based information extraction. Proc. of the 2008 IEEE 24th International Conference on Data Engineering. Washington, DC, USA. IEEE Computer Society. 2008. 933–942.

[78] Riloff E. Automatically generating extraction patterns from untagged text. Proc. of the National Conference on Artificial Intelligence (AAAI). 1996. 1044–1049.

[79] Ritter A, Clark S, Mausam, Etzioni O. Named entity recognition in tweets: An experimental study. EMNLP. ACL. 2011. 1524–1534.

[80] Romero D, Kleinberg J. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. Proc. of the 4th International AAAI Conference on Weblogs and Social Media. ICWSM '10. 2010. 138–145.

[81] Sadilek A, Kautz H, Bigham JP. Finding your friends and following them to where you are. Proc. of the Fifth ACM International Conference on Web Search and Data Mining. WSDM '12. ACM. New York, NY, USA. 2012. 723–732.

[82] Sakaguchi T, Akaho Y, Takagi T, Shintani T. Recommendations in twitter using conceptual fuzzy sets. Fuzzy Information Processing Society (NAFIPS), 2010 Annual Meeting of the North American. July 2010. 1–6.

[83] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: real-time event detection by social sensors. Proc. of the 19th International Conference on World Wide Web. WWW '10. ACM. New York, NY, USA. 2010. 851–860.

[84] Sarawagi S. Information extraction. Foundation and Trends in Databases, 2008, 1(3): 261–377.

[85] Shen W, Doan A, Naughton JF, Ramakrishnan R. Declarative information extraction using datalog with embedded extraction predicates. Proc. of the 33rd International Conference on Very Large Data Bases. VLDB '07. VLDB Endowment. 2007. 1033–1044.

[86] Shinyama Y, Sekine S. Preemptive information extraction using unrestricted relation discovery. Proc. of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Morristown, NJ, USA. Association for Computational Linguistics. 2006. 304–311.

[87] Smith M, Giraud-Carrier C. Bonding vs. bridging social capital: A case study in twitter. Social Computing (SocialCom), 2010 IEEE Second International Conference on. Aug. 2010. 385–392.

[88] Soderland S. Learning information extraction rules for semi-structured and free text. Machine Learning, February 1999, 34: 233–272.

[89] Suchanek FM, Ifrim G, Weikum G. Combining linguistic and statistical analysis to extract relations from web documents. Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '06. ACM. New York, NY, USA. 2006. 712–717.

[90] Suchanek FM, Kasneci G, Weikum G. Yago: a core of semantic knowledge unifying wordnet and wikipedia. Proc. of the 16th international conference on World Wide Web. WWW '07. ACM. New York, NY, USA. 2007. 697–706.

[91] Teutle A. Twitter: Network properties analysis. Electronics, Communications and Computer (CONIELECOMP), 2010 20th International Conference on. Feb. 2010. 180–186.

[92] Venetis P, Halve A, Madhavan J, Pasca M, Shen W, Wu F, Miao G, Wu C. Recovering semantics of tables on the web. Proc. of the VLDB Endowment. 2011.

[93] Wang RC, Cohen WW. Language-independent set expansion of named entities using the web. Proc. of the 2007 Seventh IEEE International Conference on Data Mining. Washington, DC, USA. IEEE Computer Society. 2007. 342–350.

[94] Wang RC, Cohen WW. Character-level analysis of semi-structured documents for set expansion. Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. Association for Computational Linguistics. EMNLP '09. Stroudsburg, PA, USA. 2009. 1503–1512.

[95] Welch MJ, Schonfeld U, He D, Cho J. Topical semantics of twitter links. Proc. of the Fourth ACM International Conference on Web Search and Data Mining. WSDM '11. ACM. New York, NY, USA. 2011. 327–336.

[96] Weng J, Lim EP, Jiang J, He Q. Twitterrank: finding topic-sensitive in uential twitterers. Proc. of the third ACM international conference on Web search and data mining. WSDM '10. ACM. New York, NY, USA. 2010. 261–270.

[97] Wikipedia. Wikipedia page of microblogging. http://en.wikipedia.org/wiki/microblogging.

[98] Wikipedia. Wikipedia page of twitter growth. http://en.wikipedia.org/wiki/ twitter#growth.

[99] Xu Z, Lu R, Xiang L, Yang Q. Discovering user interest on twitter with a modified author-topic model. Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on. Aug. 2011, 1: 422–429.

[100] Yardi S, Boyd D. Tweeting from the town square: Measuring geographic local networks. Proc. of the International Conference on Weblogs and Social Media. 2010. 194–201.

[101] Yates A, Cafarella M, Banko M, Etzioni O, Broadhead M, Soderland S. Textrunner: open information extraction on the web. Proc. of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. NAACL-Demonstrations '07. Association for Computational Linguistics. Stroudsburg, PA, USA. 2007. 25–26.

[102] Yin D, Hong L, Davison BD. Structural link analysis and prediction in microblogs. Proc. of the 20th ACM international conference on Information and knowledge management. CIKM '11. ACM. New York, NY, USA. 2011. 1163–1168.

[103] Yin D, Hong L, Xiong X, Davison BD. Link formation analysis in microblogs. Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11. ACM. New York, NY, USA. 2011. 1235–1236.

[104] Zhu J, Nie Z, Liu X, Zhang B, Wen JR. Statsnowball: a statistical approach to extracting entity relationships. WWW '09: Proc. of the 18th International Conference on World Wide Web. ACM. New York, NY, USA. 2009. 101–110.