Chapter 12

TEXT ANALYTICS IN SOCIAL MEDIA

Xia Hu
Computer Science and Engineering
Arizona State University
xiahu@asu.edu

Huan Liu
Computer Science and Engineering
Arizona State University
huanliu@asu.edu

Abstract The rapid growth of online social media in the form of collaboratively-created content presents new opportunities and challenges to both producers and consumers of information. With the large amount of data produced by various social media services, text analytics provides an effective way to meet usres' diverse information needs. In this chapter, we first introduce the background of traditional text analytics and the distinct aspects of textual data in social media. We next discuss the research progress of applying text analytics in social media from different perspectives, and show how to improve existing approaches to text representation in social media, using real-world examples.

Keywords: Text Analytics, Social Media, Text Representation, Time Sensitivity, Short Text, Event Detection, Collaborative Question Answering, Social Tagging, Semantic Knowledge

1. Introduction

Social media such as blogs, microblogs, discussion forums and multimedia sharing sites are increasingly used for users to communicate breaking news, participate in events, and connect to each other anytime, from anywhere. The social media sites play a very important role in current

web applications, which accounts for 50% of top 10 sites according to statistics from Alexa[1], as shown in Table 12.1. Besides that, the Twitter messages are even archived in the US Library of Congress[2]. These social media provides rich information of human interaction and collective behavior, thus attracting much attention from disciplines including sociology, business, psychology, politics, computer science, economics, and other cultural aspects of societies.

Table 12.1. Internet Traffic Report by Alexa on March 3rd, 2011

Rank	*Website*	*Rank*	*Website*
1	Google	6	**Blogger**
2	**Facebook**	7	Baidu
3	**Youtube**	8	**Wikipedia**
4	Yahoo!	9	**Twitter**
5	Windows Live	10	QQ.com

We present a definition of Social Media from a social media source, Wikipedia[3], as follows:

> "Social media are media for social interaction, using highly accessible and scalable communication techniques. It is the use of web-based and mobile technologies to turn communication into interactive dialogue."

Moturu [43] defines social media as:

> "Social Media is the use of electronic and Internet tools for the purpose of sharing and discussing information and experiences with other human beings in more efficient ways."

Traditional media such as newspaper, television and radio follow a unidirectional delivery paradigm, from business to consumer. The information is produced from media sources or advertisers and transmitted to media consumers. Different from this traditional way, web 2.0 technologies are more like consumer to consumer services. They allow users to interact and collaborate with each other in a social media dialogue of user-generated content in a virtual community. We categorize the most popular social media web sites into groups, shown in Table 12.2.

[1] www.alexa.com
[2] http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/
[3] http://en.wikipedia.org/wiki/Social_media/

Table 12.2. Types of Social Media

Category	*Representative Sites*
Wiki	Wikipedia, Scholarpedia
Blogging	Blogger, LiveJournal, WordPress
Social News	Digg, Mixx, Slashdot
Micro Blogging	Twitter, Google Buzz
Opinion & Reviews	ePinions, Yelp
Question Answering	Yahoo! Answers, Baidu Zhidao
Media Sharing	Flickr ,Youtube
Social Bookmarking	Delicious, CiteULike
Social Networking	Facebook, LinkedIn, MySpace

From the table 12.2, social media web sites contain various types of services and thus create different formats of data, including text, image, video etc. For example, the media sharing sites Flickr and Youtube allow to observe what "ordinary" users do when given the ability to more readily incorporate images and video in their everyday activity [55]. We are seeing people engaged in the creation and sharing of their personal photography. As a result, a large amount of image and video data is archived in the sites. Besides, in blogging sites, the users post frequently and create a huge number of textual / text-based data; in social bookmarking sites, users share with each other tags and URLs.

Among the various formats of data exchanged in social media, text plays a important role. The information in most social media sites (the ones with bold font in Table 12.2) are stored in text format. For example, microblogging services allow users to post small amounts of text for communicating breaking news, information sharing, and participating in events. This emerging media has become a powerful communication channel, as evidenced by many recent events like "Egyptian Revolution" and the "Tohoku earthquake and tsunami".

On the other hand, there are also a lot of useful textual data containing in the sites (the ones without bold font in Table 12.2) which are concentrating on other domains. For instance, researchers proposed to utilize tag information in multimedia sharing sites to perform video retrieval [63] and community detection [59]. Under these scenarios, how to mine useful information from textual data presents great opportunities to social media research and applications.

Text Analytics (also as know as Text Mining) refers to the discovery of knowledge that can be found in text archives [49]. This field has received much attention due to its wide application as a multi-purpose tool, borrowing techniques from Natural Language Processing (NLP),

Data Mining (DM), Machine Learning (ML), Information Retrieval (IR) etc.

Text Analytics is defined in Wikipedia as follows:

> "Text Analytics describes a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation."

Text analytics techniques can help efficiently deal with textual data in social media for research and business purposes. The rest of this chapter is organized as follows: Section 2 introduces specialty for text analytics in social media by analyzing the features of textual data. Section 3 presents proposed approaches for several representative research issues. Section 4 introduces one example to illustrate in detail the process of text analytics methods to solve real world problems. Section 5 concludes the chapter with some possible directions of future work.

2. Distinct Aspects of Text in Social Media

Textual data in social media gives us insights into social networks and groups that were not previously possible in both scale and extent. Unfortunately, textual data in social media presents many new challenges due to its distinct characteristics. In this section, we first review traditional processes of text analytics and then discuss the distinctive features of text in social media, including *Time Sensitivity, Short Length, Unstructured Phrases, Abundant Information.*

2.1 A General Framework for Text Analytics

In this subsection, we briefly introduce the general framework of text analytics to process a text corpus. A traditional text analytics framework consists of three consecutive phases: Text Preprocessing, Text Representation and Knowledge Discovery, shown in Figure 12.1. We use an example to illustrate these methods in each step.

Given a text corpus which contains three microblogging messages, as shown below:

"watching the King's Speech"
"I like the King's Speech"
"they decide to watch a movie"

Text Preprocessing: Text preprocessing aims to make the input documents more consistent to facilitate text representation, which is

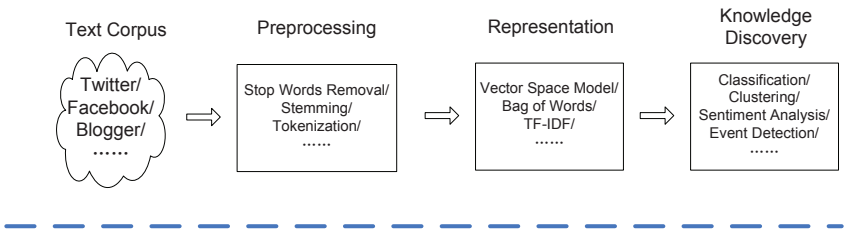

Figure 12.1. A Traditional Framework for Text Analytics

necessary for most text analytics tasks. Traditional text preprocessing methods include *stop word removal* and *stemming*. Stop word removal eliminates words using a stop word list[4], in which the words are considered more general and meaningless; Stemming [46] reduces inflected (or sometimes derived) words to their stem, base or root form. For example, "watch", "watching", "watched" are represented as "watch", so the words with variant forms can be regarded as same feature. The output of text preprocessing for the three microblogging messages are:

"watch King' Speech"
"King' Speech"
"decid watch movi"

Preprocessing methods depend on specific application. In many applications, such as Opinion Mining or NLP, they need to analyze the message from a syntactical point of view, which requires that the method retains the original sentence structure. Without this information, it is difficult to distinguish "Which university did the president graduate from?" and "Which president is a graduate of Harvard University?", which have overlapping vocabularies. In this case, we need to avoid removing the syntax-containing words.

Text Representation: The most common way to model documents is to transform them into sparse numeric vectors and then deal with them with linear algebraic operations. This representation is called "Bag Of Words" (BOW) or "Vector Space Model" (VSM). In these basic text representation models, the linguistic structure within the text is ignored and thus leads to "structural curse".

In BOW model, a word is represented as a separate variable having numeric weight of varying importance. The most popular weighting

[4]http://www.lextek.com/manuals/onix/

schema is Term Frequency / Inverse Document Frequency (TF-IDF):

$$tfidf(w) = tf * log\frac{N}{df(w)}, \tag{12.1}$$

where:

– tf(w) is term frequency (the number of word occurrences in a document)

– df(w) is document frequency (the number of documents containing the word)

– N is number of documents in the corpus

– tfidf(w) is the relative weight of the feature in the vector

Using BOW to model the three messages with a TF-IDF weight, the corpus can be represented as a words * documents matrix. Each row represents a word (5 distinct words in total) and each column represents a message, as shown below:

$$\begin{bmatrix} watch \\ King' \\ Speech \\ decid \\ movi \end{bmatrix} = \begin{bmatrix} 0.4055 & 0 & 0.4055 \\ 0.4055 & 0.4055 & 0 \\ 0.4055 & 0.4055 & 0 \\ 0 & 0 & 1.0986 \\ 0 & 0 & 1.0986 \end{bmatrix} \tag{12.2}$$

Knowledge Discovery: When we successfully transform the text corpus into numeric vectors, we can apply the existing machine learning or data mining methods like classification or clustering. For example, in machine learning, similarity is an important measure for many tasks. A widely used similarity measure between two messages V_1 and V_2 is cosine similarity, which can be computed as:

$$similarity(V_1, V_2) = cos(\theta) = \frac{V_1 * V_2}{||V_1||||V2||}, \tag{12.3}$$

By conducting text preprocessing, text representation and knowledge discovery methods, we can mine latent, useful information from the input text corpus, like similarity between two messages in our example. However, this presents challenges for traditional text analytics methods when applied directly to textual data in social media due to its distinct features. Now we analyze the new features of textual data in social media from four different perspectives: Time Sensitivity, Short Length, Unstructured Phrases, and Abundant Information.

2.2 Time Sensitivity

An important and common feature of many social media services is their real-time nature. Particularly, bloggers typically update their blogs

every several days, while microblogging and social networking users may post news and information several times daily. Users may want to communicate instantly with friends about "What are you doing?" (Twitter) or "What is on your mind" (Facebook). When submitting a query to Twitter, the returned results are only several minutes old.

Besides communicating and sharing minds with each other, users post comments on recent events, such as new products, movies, sports, games, political campaigns, etc. The large number of real-time updates contain abundant information, which provides a lot of opportunities for detection and monitoring of an event. With these data, we are able to infer a user's interest in an event [37], and track information provenance from the user's communications [9]. For example, Sakaki et al. [47] investigate the real-time interaction of events such as earthquakes, and they propose an algorithm to monitor tweets and to detect a target event.

With the rapid evolution of content and communication styles in social media, text is changing too. Different from traditional textual data, the text in social media is not independent and identically distributed (i.i.d.) data anymore. A comment or post may reflect the user's interest, and a user is connected and influenced by his friends. People will not be interested in a movie after several months, while they may be interested in another movie released several years ago because of the recommendation from his friends; reviews of a product may change significantly after some issues, like the comments on Toyota vehicles after the break problem. All these problems originate from the time sensitivity of textual data in social media.

2.3 Short Length

Certain social media web sites restrict the length of user-created content such as microblogging messages, product reviews, QA passages and image captions, etc. Twitter allows users to post news quickly and the length of each tweet is limited to 140 characters. Similarly, Picasa comments are limited to 512 characters, and personal status messages on Windows Live Messenger are restricted to 128 characters. As we can see, data with a short length is ubiquitous on the web at present. As a result, these short messages have played increasing important roles in applications of social media. Successful processing short texts is essential to text analytics methods.

Short messages, as the most important data format, make people more efficient with their participate in social media applications. However, this brings new challenges to traditional fundamental research topics in text analytics, such as text clustering, text classification, infor-

mation extraction and sentiment analysis. Unlike standard text with lots of words and their resulting statistics, short messages consist of few phrases or sentences. They cannot provide sufficient context information for effective similarity measure [45], the basis of many text processing methods [27].

To tackle the data sparseness problem, several traditional text analytics methods have been proposed, which can be generally categorized into two groups. The first is the basic representation of texts called surface representation [32, 36], which exploits phrases in the original text from different aspects to preserve the contextual information. However, NLP techniques such as parsing are not employed, as it is time consuming to apply such techniques to analyze the structure of standard text in detail. As a result, the methods fail to perform a deep understanding of the original text. Another limitation of such methods is that they did not use external knowledge, which has been found to be useful in dealing with the semantic gap in text representation [18]. For example, tag "Japan Earthquake" does not contain any words or phrases related to "Nuclear Crisis" while we learn that these two events are related from recent news. Because they have no common words or phrases, it is very difficult for BOW-based models and methods to build semantic connections between each other. One intuitive approach is to enrich the contexts of basic text segments by exploiting external resources, and such methods have been found to be effective in narrowing the *semantic gap* in different tasks [20, 54].

2.4 Unstructured Phrases

An important difference between the text in social media and traditional media is the variance in the quality of the content. First, the variance of quality originates from people's attitudes when posting a microblogging message or answering a question in a forum. Some users are experts for the topic and post information very carefully, while others do not post as high of quality. The main challenge posed by content in social media sites is the fact that the distribution of quality has high variance: from very high-quality items to low-quality, sometimes abusive content. This makes the tasks of filtering and ranking in such systems more complex than in other domains [5].

Second, when composing a message, users may use or coin new abbreviations or acronyms that seldom appear in conventional text documents. For example, messages like "How r u?", "Good 9t" are not really words, but they are intuitive and popular in social media. They provide users convenience in communicating with each other, however it is very

difficult to accurately identify the semantic meaning of these messages. Besides the unstructured expressions, the text is sometimes "noisy" for a specific topic. For instance, one QA passage in Yahoo! Answers "I like sony" should be noisy data to a post that is talking about iPad 2 release. It is difficult to classify the passage into corresponding classes without considering its context information.

2.5 Abundant Information

Social media in general exhibit a rich variety of information sources. In addition to the content itself, there is a wide array of non-content information available. For example, Twitter allows users to utilize the "#" symbol, called hashtag, to mark keywords or topics in a Tweet (tag information); an image is usually associated with multiple labels which are characterized by different regions in the image [66]; users are able to build connection with others (link information) in Facebook and other social network sites; Wikipedia provides an efficient way for users to redirect to the ambiguity concept page or higher level concept page (semantic hierarchy information).

All these external information presents opportunities for traditional tasks. Previous text analytics sources always appear as <user, content> structure, while the text analytics in social media is able to derive data from various aspects, which include user, content, link, tag, time stamp etc. Recently, many research work utilizes link information in microblogging services to detect the popular event [37], distinguish the microblogging message is credible news or just rumor [42]. Also, with the user metadata (e.g. tags) mined from blogosphere and bookmarking sites, Wang et al. [59] take advantage of networking information between users and tags to discover overlapping communities. These successful applications motivated us to exploit more opportunities behind such abundant additional information available in social media.

3. Applying Text Analytics to Social Media

It presents great challenges to apply traditional methods to process textual data in social media. Recently, a number of methods have been proposed to handle the textual data with new features. In this section, we introduce a variety of applying text analytics to social media.

3.1 Event Detection

Event Detection aims to monitor a data source and detect the occurrence of an event that is captured within that source [40]. These data sources include images, video, audio, spatio-temporal data, text docu-

ments and relational data. Among them, event detection and evolution tracking of news articles [60], digital books [22] receives much attention. The volume of textual data in social media is increasing exponentially, thus providing us many opportunities for event detection and tracking.

In some sense, social text streams are sensors of the real world [67]. As the real-time nature of textual data in social media, a lot of work has been done to extract real world events from social text streams. One interesting application is to monitor real-time events. For example, when an earthquake or tsunami occurs, one convenient way to communicate updated news with others is to post messages related to the event via microblogging. Therefore, it provides possibility for us to promptly detect the occurrence of earthquake or tsunami, simply by mining the corresponding microblogging messages. Based on the above observation, Sakaki et al. [47] investigate the real-time interaction of events on Twitter. They consider each user as a sensor to monitor tweets posted recently and to detect earthquake or rainbow. To detect a target event, the work flow is as follows. First, a classifier is trained by using keywords, message length, and corresponding context as features to classify tweets into positive or negative cases. Second, they build a probabilistic spatio-temporal model for the target event to identify location of the event. As an application, the authors constructed an earthquake-reporting system in Japan, where has numerous earthquakes every year as well as a large number of active microblogging users.

One important direction of event detection in social media is to improve traditional news detection. A large number of news stories are generated from various news channels day after day. Among them, only a relatively few receive attention from users, which are recognized as "breaking news". Traditionally, editors of newspapers and websites decide which stories can be ranked higher and assigned in an important place, like the front page. In a similar way, web-based news aggregated services, such as Google News[5], give users access to broad perspectives on the important news stories being reported by grouping articles into related news events. Deciding automatically on which top stories to show is a challenging problem [39]. A poll conducted by Technorati found that 30% of bloggers consider themselves to be blogging about news-related topics [41].

Motivated by this observation, researchers proposed to utilize blogosphere to facilitate news detection and evaluation. Lee et al. present novel approaches to identify important news story headlines from the

[5]http://news.google.com/

blogosphere for a given day [34]. The proposed system consists of two components based on the language model framework, the query likelihood and the news headline prior. For the query likelihood, the authors propose several approaches to estimate the query language model and the news headline language model. They also suggest several criteria to evaluate the importance or newsworthiness of the news headline for a given day.

Tracking the diffusion and evolution of a popular event in social media is another interesting direction in this field. Different from i.i.d. textual data in traditional media, user generated content in social media is a mixture of a text stream and a network structure. Lin et al. take into account the burstiness of user interest, information diffusion in the network structure and the evolution of textual topics to model the popularity of events over time [37]. They tackle the problem of popular event tracking in online communities by studying the interplay between textual content and social networks.

Besides detecting events from pure textual data, some methods have been proposed to mine text information in social media to facilitate event detection. Chen et al. is to detect events from photos on Flickr by analyzing the tag of the photos [13]. In the proposed framework, the authors first analyze temporal and locational distributions of tag usage. Second, they identify tags related with events, and further distinguish if the tags are relevant to aperiodic events or periodic events. Afterwards, tags are clustered into their corresponding clusters. Each cluster represents an event, and consists of tags with similar temporal and locational distribution patterns as well as with similar associated photos. Finally, for each tag cluster, photos corresponding to the represented event are extracted.

3.2 Collaborative Question Answering

Collaborative question answering services begin to emerge with the blooming of social media. They bring together a network of self-declared "experts" to answer questions posted by other people. A large volume of questions are asked and answered every day on social Question and Answering (QA) web sites such as Yahoo! Answers. Collaborative question answering portals are a popular destination for users looking for advice with a particular situation, for gathering opinions, for sharing technical knowledge, for entertainment, for community interaction, and for satisfying one's curiosity about a countless number of things.

Over time, a tremendous amount of historical QA pairs have built up their databases, and this transformation gives users an alternative

place to look for information, as opposed to a web search. Instead of looking through a list of potentially relevant documents from the Web or posting a new question in a forum, users may directly search for relevant historical questions or answers from QA archives. As a result, the corresponding best solutions could be explicitly extracted and returned.

This problem could be considered from two sides. On one hand, the most relevant questions semantically related to the query are returned, so that users can find similar questions and their corresponding answers. Wang et al. [57] propose a graph based approach to perform question retrieval by segmenting multi-sentence questions. The authors first attempt to detect question sentences using a classifier built from both lexical and syntactic features, and use similarity and co-reference chain based methods to measure the closeness score between the question and context sentences. On the other hand, systems provide corresponding quality QA pairs from answer's point of view. Adamic et al. [1] evaluate the quality of answers for specific question by analyzing Yahoo! Answer's knowledge sharing activity. First, forum categories are clustered according to the content characteristics and patterns of interaction among users. The interactions in different categories reveal different characteristics. Some categories are more like expertise sharing forums, while others incorporate discussion, everyday advice, and support. Similarly, some users focus narrowly on specific topics, while others participate across categories. Second, the authors utilize this feature to map related categories and characterize the entropy of the users' interests. Both user attributes and answer characteristics are combined to predict, within a given category, whether a particular answer will be chosen as the best answer by the asker.

In order to improve QA archives management, there are a number of works done by evaluating the quality of QA pairs. Harper et al. [25] tried to determine which questions and answers have archival value by analyzing the differences between conversational questions and informational questions. Informational questions refer to the questions with the intent of obtaining information the asker could learn from. An example is "Is drinking Coke good for health?". Conversational questions refer to the questions with the intent of stimulating discussion. In these questions, the users may aim at getting opinions or self-expression. An example is "Do you like drinking Coke?". The authors present evidence that conversational questions typically have much lower potential archival value than informational questions. Further, they used machine learning techniques to automatically classify questions as conversational or informational from perspectives of the process about categorical, linguistic, and social differences between different question types. Agichtein et al. [5]

introduced a general classification framework for combining the evidence from different sources, that can be tuned automatically for quality prediction of QA pairs. In particular, they exploit features of QA pairs that are intuitively correlated with quality, including intrinsic content quality, interactions between content creators and users, as well as the content usage statistics. Then a classifier is trained to appropriately select and weight the features for each specific type of item, task, and quality definition.

3.3 Social Tagging

Social tagging is a method for Internet users to organize, store, manage and search for tags / bookmarks (also as known as social bookmarking) of resources online. Unlike file sharing, the resources themselves aren't shared, merely the tags that describes them or bookmarks that reference them[6]. The rise of social tagging services presents a potential great deal of data for mining useful information on the web. The users of tagging services have created a large volume of tagging data which has attracted recent attention from the research community. From oceans of tags, it is difficult for a user to quickly locate the relevant resources he wants via browsing the tags. Typically, the tagging services provide keyword-based search which returns resources annotated by the given tags. However, the results returned by the search module are inadequate for users to discover interesting resources due to the short and unstructured nature of tags. First, it is very difficult to design an effective tag ranking algorithm due to the short length and sparseness of tags. Second, current systems are designed for keywords based search, which failed to capture the semantic relationship between two semantically related tags. For example, when a user searches for a recent event, such as "Egyptian Revolution", the systems will return results that are tagged as "Egyptian" or "Revolution". Among them, resources tagged with "Mubarak's resignation" and "Protest" which are highly related to "Egyptian Revolution" will be ignored. This "semantic gap" results in many valuable and interesting results overlooked and buried in disorganized resources.

Research work in social tagging services can be typically divided into two categories: one aims to improve the quality of tag recommendation and the other studies how to utilize social tagging resources to facilitate other applications. First, Sigurbjornsson and Van [48] investigate how to assist users during the tagging phase in multimedia sharing sites (Flickr).

[6] http://en.wikipedia.org/wiki/Social_bookmarking/

They present and evaluate tag recommendation strategies to support the user in the photo annotation task by recommending a set of tags that can be added to the photo. Yin et al. [61] address the problem of tag prediction by proposing a probabilistic model for personalized tag prediction. On the other hand, social tagging resources are exploited to improve other web applications, including web object classification [62], document recommendation [23], web search quality [26] etc.

3.4 Bridging the Semantic Gap

As we discussed in Section 2, the textual data in social media is short and unstructured. When processing this kind of data, traditional bag of words (BOW) approach is inherently limited, as it can only use pieces of information that are explicitly mentioned in the documents [18]. Consider one famous movie "The Dark Knight". By mining the original posts related to this movie, it is inadequate to build the semantic relationship with other relevant concepts due to the *semantic gap*. For example, "The Dark Knight" and "Batman" are different names of one movie, but they cannot be linked as the same concept without additional information from external knowledge. Specifically, this approach has no access to the wealth of world knowledge possessed by humans, and is easily puzzled by facts and terms not mentioned in the data set. Recently, researchers have proposed semantic knowledge bases to bridge the widely extant semantic gap in short text representation.

The aggregation of information in groups is often better than what could have been made by any single member of the group [52]. Wikipedia is a free, web-based, collaborative, multilingual encyclopedia project. Its 18 million articles have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone with access to the site[7]. Unlike other standard ontologies, such as WordNet or Mesh, Wikipedia is not a structured thesaurus edited by experts, but it was contributed collaboratively by users on the web. It is comprehensive, up to date and well-formed [29]. In Wikipedia, each article concentrates on one specific topic. The title of each article is a succinct phrase that resembles an ontology term. Equivalent concepts are grouped together by redirected links. Meanwhile, Wikipedia contains a hierarchical categorization system, in which each article belongs to at least one category. All these features are making Wikipedia a potential ontology for enhancing text representation.

[7] http://en.wikipedia.org/wiki/Wikipedia/

Some methods were proposed to tackle the problems of data sparseness and the semantic gap in short texts clustering and classification by exploiting semantic knowledge. Somnath et al. [8] proposed a method to enrich short text representation with additional features from Wikipedia. The method used titles of Wikipedia articles as additional external features, and it showed improvement in the accuracy of short texts clustering. Phan et al. [45] presented a framework for building classifiers that deal with short texts from the Web and achieved qualitative enhancement. The underlying idea of the framework is to collect large-scale data and then build a classifier on both labeled data and external semantics for each classification task. In addition, researchers [56, 18, 58] analyzed the documents and found related ontological concepts within WordNet and Wikipedia, in turn producing a set of features that augment standard BOW. Towards improving the management of Google snippets, existing methods focus either on classifying the web texts into smaller categories [28] or assigning labels for each category [10] with the help of Wikipedia.

3.5 Exploiting the Power of Abundant Information

Abundant information associated with textual information is ubiquitous in social media. On Twitter, for example, two microblogging messages can be linked together via their authors' follower, followee, retweet or reply relationship; two microblogging messages can be classified into the same or similar category when they share the same hashtag or contain same hyperlink; semantic similarity between two microblogging messages can be measured based on their posting time (time stamp), posting place (geotag), author's personal information (profile), etc. Similar phenomena can be observed in Facebook, LinkedIn, Wikipedia and other social media sites. Different from i.i.d. documents in traditional media, if one can utilize these abundant information available in social media, performance of many text analytics methods may be significantly improved.

To utilize the abundant information appearing along with text content in social media, recent methods have been proposed to integrate this into text analytics tasks, including classification, clustering etc.. Among these methods, a combination of link and text content for mining purposes is becoming popular. A major difference between these two kinds of methods is that traditional methods measure the similarity between documents purely based on attribute similarity (e.g. cosine similarity between two attribute vectors); while the methods for text in social media

measures document similarity based on connectivity (e.g. the number of possible paths between authors of the documents) and structural similarity (e.g. the number of shared neighbors) [68], besides the attribute similarity of text content. Links clearly contain high-quality semantical clues that are lost in purely text-based methods, but exploiting link information is not easy. The major difficulty is the similarity measurement between each pair of objects, due to the characteristics of differing social networks:

- *Multi-dimensional social networks.* The connections between users in social media are often multi-dimensional [53]. Users can connect to each other for different reasons, e.g., alumni, colleagues, living in the same community, sharing similar interests, etc. Different types of links have different semantic meanings associated with their respective latent social dimensions.

- *Network representation.* Traditional text analytics methods utilize local features or attributes to represent documents. However, there is no natural feature representation for all types of network data [31]. When we use an adjacency matrix to represent a network, the matrix will be very sparse, highly dimensional and its equal weights cannot reflect tie strength well. Moreover, obtaining labels of objects in social network, which is very important for supervised learning methods, appears to be very expensive.

- *Dynamic networks.* Different from constant news collections or a documents corpus, social media is evolving continuously, with new users joining the network, extant users connecting with each other or becoming dormant. It is imperative to update the acquired community structure. As a result, how to efficiently integrate the updated network information is very important for many applications.

Many methods have been proposed to tackle the challenges and make use of link information sources. To our knowledge, the first topic classification system that simultaneously utilizes textual and link features was discussed in [11]. The authors aim to propose a statistical model and a relaxation labeling technique to build a classifier by exploiting link information form neighbors of the documents. Similarly Furnkranz [17] found that it is possible to classify documents more reliably with information originating from pages that point to the document than with features that are derived from the document itself. Later, Chakrabarti et al. [6] proposed a graph-based text classification method by learning

from their neighbors. The difference between these two kinds of techniques is that the latter one considered more factors in social networks, including the network evolution (dynamic network), pruning of edges from the neighborhood graph, and weighing the influence of edges and edges themselves by content similarity measures. Recently, Aggarwal and Li [3] presented an efficient and scalable method to tackle the problem of node classification in dynamic information networks with both text content and links. To facilitate an effective classification process, different from previous models, they use a random walk approach in conjunction with the content of the network. This design makes the model more robust to variations in content and linkage structure. Aside from classification, link information has been also successfully integrated into the applications of clustering [68] and topic modeling [50]. It shows that the use of both link and text information achieved more effective results than a method based purely on either of the two [4].

In addition to integrating network information into text analytics tasks, researchers further exploit abundant information. In [51], a heterogeneous information network is defined as an information network composed of multiple types of object. The authors explored clustering of multi-typed heterogeneous networks with a star network schema, although clustering on homogeneous networks has been well studied over decades. Links across multi-typed objects are utilized to generate high-quality net-clusters. The general idea of the proposed framework is to avoid measuring the pairwise similarity between objects, which is hard in heterogeneous networks.Instead, it maps each target object into a low dimensional space defined by current clustering results. Then every target object in these clusters will be readjusted based on the new measure. The clustering results will be improved in each iteration until convergence.

3.6 Related Efforts

Aside from the topics discussed in the previous sections, even more attempts have been explored in mining social media resources. In Social Network Analysis, researchers utilize various information such as the posts, links, tags, etc., to identify influential users in the blogsphere [2] and microblogsphere [7], to understand user behavior in microblogsphere by analyzing the user intentions associated at a community level [30, 33]. In Sentiment Analysis, Conner et al. investigate several surveys on consumer confidence and political opinion, connect measures of public opinion measured from polls with sentiment measured from text [14]. Gerani et al. use a general opinion lexicon and propose using proximity

information in order to capture opinion term relatedness to improve opinion retrieval in the blogsphere [21]. In Knowledge Management, Lerman and Hogg [35] use a model of social dynamics to predict the popularity of news. Incorporating aspects of web site design, the model improves on predictions based on simply extrapolating from early votes. Lu et al. exploit contextual information about the authors' identities and social networks for improving review quality prediction [38]. This model improves previous work, which addressed the problem by treating a review as a stand-alone document, extracting features from the review text, and learning a function based on these features for predicting the review quality.

4. An Illustrative Example

In this section, we present one real world application to further illustrate how to utilize text analytics to solve problems in social media applications. We now introduce an effective way to improve the short text representation quality by integrating semantic knowledge resources.

As we discussed in Section 2, textual data in social media has the problems of data sparseness and semantic gap. One effective way to solve these problems is to integrate semantic knowledge, which has been found to be useful in dealing with the semantic gap [18]. For example, the first search result returned by Google using "Friday" as the query does not contain any words or phrases related to "Rebecca Black ", while we learn that the singer creates overnight sensations by sharing the song via YouTube. Because they have no words or phrases overlapping, this result can not be successfully build connection with Rebecca related content. Thus, one intuitive idea is to enrich the contexts of basic text representation by exploiting semantic resources.

Now, we follow the basic idea proposed in [28] to illustrate three steps of feature generation in detail: Seed Phrase Extraction from the original text corpus, Semantic Features Generation based on *seed phrases* and Feature Space Construction.

4.1 Seed Phrase Extraction

Given a text corpus, features can be derived by employing different techniques in NLP. The only requirement is that the extracted features could be informative to cover the key subtopics described in the short texts. Here we use shallow parsing [24] to divide sentences into a series of words that together compose a grammatical unit. To ensure the extracted features are able to cover main topics, we use these phrases generated by shallow parser, with the combination of sentences in the

original text, to extract the *seed phrases*. However, there are redundancies between these two kinds of features. If we employ all these features as *seed phrases*, they would produce some duplicate information between each other. Therefore, to make the tradeoff between informativeness and effectiveness, we propose to measure the semantic similarity between sentence level features and phrase level features to eliminate information redundancy.

Several methods have been proposed to calculate the semantic similarity between associations [12] using web search. However, along with the increasing scale of the web, the page counts provided by some commercial search engines are not so reliable [15]. Thus instead of simply using the search engine page counts, we propose a phrase-phrase semantic similarity measuring algorithm using a co-occurrence double check in Wikipedia to reduce the semantic duplicates. For Wikipedia, we download the XML corpus [16], remove xml tags and create a Solr [8] index of all XML articles.

Let T denote a *sentence level* feature, $T = \{t_1, t_2, \ldots, t_n\}$, where t_i denotes the *phrase level* feature contained in T. The *sentence level* feature is too sparse to calculate its frequency directly. Therefore, we calculate the semantic similarity between t_i and $\{t_1, t_2, \ldots, t_n\}$ as $InfoScore(t_i)$ instead. We select the *phrase level* feature which has the largest similarity with other features in T and remove it as the redundant feature.

Given two phrases t_i and t_j, we use t_i and t_j separately as a query to retrieve top C Wikipedia pages from the built index. The total occurrences of t_i in the top C Wikipedia pages retrieved by query t_j is denoted as $f(t_i|t_j)$; and we define $f(t_j|t_i)$ in a similar manner. The total occurrences of t_i in the top C Wikipedia pages retrieved by query t_i is denoted as $f(t_i)$, and similarly for $f(t_j)$. The variants of three popular co-occurrence measures [15] are defined as below:

$$WikiDice(t_i, t_j) = \begin{cases} 0 & \text{if } f(t_i \mid t_j) = 0 \\ & \text{or } f(t_j \mid t_i) = 0 \\ \frac{f(t_i|t_j)+f(t_j|t_i)}{f(t_i)+f(t_j)} & \text{otherwise} \end{cases}, \tag{12.4}$$

where WikiDice is a variant of the Dice coefficient.

$$WikiJaccard(t_i, t_j) = \frac{min(f(t_i \mid t_j), f(t_j \mid t_i))}{f(t_i) + f(t_j) - max(f(t_i \mid t_j), f(t_j \mid t_i))}, \tag{12.5}$$

where WikiJaccard is a variant of the Jaccard coefficient.

[8] http://lucene.apache.org/solr/

$$WikiOverlap(t_i, t_j) = \frac{min(f(t_i \mid t_j), f(t_j \mid t_i))}{min(f(t_i), f(t_j))}, \tag{12.6}$$

where WikiOverlap is a variant of the Overlap(Simpson) coefficient.

For ease of comparison, all the $\frac{n^2}{2}$ WikiDice similarity scores are normalized into values in $[0, 1]$ range using the linear normalization formula defined below:

$$WD_{ij} = \frac{WikiDice_{ij} - min(WikiDice_k)}{max(WikiDice_k) - min(WikiDice_k)}, \tag{12.7}$$

where k is from 1 to $\frac{n^2}{2}$. We again define WJ_{ij} and WO_{ij} in a similar manner. A linear combination is then used to incorporate the three similarity measures into an overall semantic similarity between two phrases t_i and t_j, as follows:

$$WikiSem(t_i, t_j) = (1 - \alpha - \beta)WD_{ij} + \alpha WJ_{ij} + \beta WO_{ij}, \tag{12.8}$$

where α and β weight the importance of the three similarity measures. Text clustering is an unsupervised method where we do not have any labeled data to tune the parameters. We thus empirically set α and β to equal weight.

For each *sentence level* feature, we rank the information score defined in Equation 12.9 for its child node features at *phrase level*.

$$InfoScore(t_i) = \sum_{j=1, j \neq i}^{n} WikiSem(t_i, t_j). \tag{12.9}$$

Finally, we remove the *phrase level* feature t^*, which delegates the most information duplicates to the *sentence level* feature T, defined as:

$$t^* = \arg \max_{t_i \in \{t_1, t_2, \ldots, t_n\}} InfoScore(t_i). \tag{12.10}$$

4.2 Semantic Feature Generation

After extracting the *seed phrases* from the first step, we obtain an informative and effective basic representation of the input text corpus. In this step, we discuss an algorithm to generate semantic features based on the *seed phrases* using Wikipedia as background knowledge.

4.2.1 Background Knowledge Base. Wikipedia, as background knowledge, has a wider knowledge coverage than other semantic knowledge bases and is regularly updated to reflect recent events. Under this scenario, we take Wikipedia as the semantic knowledge source to generate semantic concepts.

Gabrilovich and Markovitch [19], as well as Hu et al. [27] preprocessed the Wikipedia corpus to collect semantic concepts. Preprocessing

Algorithm 1: GenerateFeatures(S)

input : a set S of *seed phrases*
output: *semantic features* SF

SF $\leftarrow$ *null*
for *seed phrase* $s \in S$ **do**
  if $s \in$ Sentence level **then**
    s.Query $\leftarrow$ SolrSyntax(s, OR)
  else
    s.Query $\leftarrow$ SolrSyntax(s, AND)
  WikiPages $\leftarrow$ Retrive(s.Query)
  $SF \leftarrow SF$ + Analyze(WikiPages)
return SF

Figure 12.2. Semantic feature generation scheme

Wikipedia is one way to build the concepts space. However, it ignores the valuable contextual information of Wikipedia plain texts and always encounters problems when mapping the original text to appropriate concepts. Therefore, in this study we introduce another way to process the Wikipedia corpus, it is to preserve the original pages of Wikipedia with the built-in Solr index.

4.2.2 Feature Generator. The *semantic feature* generation algorithm is illustrated in Figure 12.2. Given a *seed phrase*, we retrieve corresponding Wikipedia pages with the help of the Solr search engine. Then we extract semantic concepts from the retrieved Wikipedia pages.

In order to retrieve the appropriate pages from the large Wikipedia corpus, we derive queries based on *seed phrase* arising from *sentence level* or *phrase level* separately. As the key information of *seed phrases* from *phrase level* is more focused, we build the "AND" query which requires the retrieved pages to contain every term in the phrase. On the other hand, the *seed phrases* from *sentence level* are informative but sparse, we thus build "OR" query[9] which means there is no guarantee that the retrieved Wikipedia pages will contain every term in the phrase. We use these two kinds of queries to retrieve the top ω articles from the Wikipedia corpora. Similar to [8], we extract titles and bold terms (links) from the retrieved Wikipedia pages to serve as part

[9]For more details about "AND" and "OR" query syntax, please refer to http://wiki.apache.org/solr /SolrQuerySyntax/

of the *semantic features*. To discover the intrinsic concepts hidden in the plain texts, we adopt an effective key phrase extraction algorithm — Lingo [44], which uses algebraic transformations of the term-document matrix and implements frequent phrase extraction using suffix arrays. The key phrases extracted from the Wikipedia pages are added to the *semantic feature* space. By utilizing this method, we may obtain extrinsic concepts "Friday" for the phrase "Rebecca Black" and the intrinsic concepts like "Song", "Singer" and "Youtube" by mining the related pages. Therefore, we can build semantic relationships between the concepts of"Friday" and "Rebecca Black".

4.3 Feature Space Construction

As the construction of Wikipedia follows the non-binding guidelines and the data quality is only under social control by the community [65], it often introduces noise to the corpus. Meanwhile, a single text may generate a huge number of features. These overzealous *semantic features* bring adverse impact on the effectiveness and dilute the influence of valuable original information. Therefore, we conduct feature filtering to refine the unstructured or meaningless features and apply feature selection to avoid aggravating the "curse of dimensionality".

Feature Filtering: We formulate empirical rules to refine the unstructured features obtained from Wikipedia pages, some typical rules are as follows:

- Remove features generated from too general *seed phrase* that returns a large number (more than 10,000) of articles from the index corpus.
- Transform features used for Wikipedia management or administration, e.g. "List of hotels"→"hotels", "List of twins"→"twins".
- Apply phrase sense stemming using Porter stemmer [46], e.g. "fictional books"→"fiction book".
- Remove features related to chronology, e.g. "year", "decade" and "centuries".

Feature Selection: We need to select *semantic features* to construct feature space for various tasks. The number of *semantic features* we need to collect is determined by the specific task. In this chapter, we utilize a simple way to select the most frequent features.

First, the *tf-idf* weights of all generated features are calculated. One *seed phrase* $s_i (0 < i \leq m)$ may generate k *semantic features*, denoted by $\{f_{i1}, f_{i2}, \ldots, f_{ik}\}$. In order to explore the diversity of the *semantic*

features, we select one feature for each *seed phrase*. Thus m features are collected as follows:

$$f_i^* = \arg \max_{f_{ij} \in \{f_{i1}, f_{i2}, \ldots, f_{ik}\}} tf\_idf(f_{ij}). \tag{12.11}$$

Second, the top n features are extracted from the remaining *semantic features* based on their frequency. These frequently appearing features, together with the features from the first step, are used to construct the $m + n$ *semantic features*.

Now we prepare the feature space for clustering, classification or other text analytics methods. From the discussion above, key idea of the framework is to introduce semantic knowledge base (Wikipedia) to build semantic connection between two short documents. This section provides a clear mind about how to apply text analytics methods in social media resources.

5. Conclusion and Future Work

Textual data in social media carries abundant information. User-generated content provides diverse and unique information in forms of comments, posts and tags. The useful information hidden in the text resources of social media provides opportunities for researchers of different disciplines to mine patterns and information of interest that might not be obvious. In this chapter, we discuss about the distinct aspects of textual data in social media and their challenges, and elaborate current work of utilizing text analytics methods to solve problems in social media.

This chapter has only discussed some essential issues. There are a number of interesting directions for further exploration.

- How to better make use of the real-time nature in social media? A real-time search system which can find, summarize and track updated breaking news or events in social communities will be very challenging but useful.

- How to handle textual data with short length in social media? As we discussed, short text plays a very important role in social media. On one hand, these textual data contains less information than standard documents; on the other hand, it provides possibility for us to use traditional syntax-based NLP models to perform fine-level textual analysis, which were very time consuming for standard text.

- How to exploit cross media data to facilitate social behavior analysis? Cross media data here refers data of different formats or data from different social media resources [64]. The variance types of data in social media, including text, image, link or even multilingual data, have latent relationships and interactions between each other. Also, an efficient and effective way to integrate these kinds of data will be very useful to address the data sparseness problem.
- How to process web scale data available in social media? The large volume and the compact but noisy presentation of textual data in social media hinders the accessibility of information for users to conveniently search, navigate and locate the specific messages one might be interested in. Finding an efficient way to handle these large scale data types is very challenging.

Acknowledgments

This work is, in part, supported by the grants NSF (#0812551), ONR (N000141010091) and AFOSR (FA95500810132). The authors would like to acknowledge all of the researchers in Arizona State University's Data Mining and Machine Learning Laboratory. The views expressed in this chapter are solely attributed to the authors and do not represent the opinions or policies of any of the funding agencies.

References

[1] L. Adamic, J. Zhang, E. Bakshy, and M. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceeding of the 17th international conference on World Wide Web*, pages 665–674. ACM, 2008.

[2] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 207–218, New York, NY, USA, 2008. ACM.

[3] C. C. Aggarwal and N. Li. On node classification in dynamic content-based networks. In *The Eleventh SIAM International Conference on Data Mining*, pages 355–366, 2011.

[4] C. C. Aggarwal and H. Wang. Text mining in social networks. *Social Network Data Analytics*, pages 353–378, 2011.

[5] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 183–194, New York, NY, USA, 2008. ACM.

[6] R. Angelova and G. Weikum. Graph-based text classification: learn from your neighbors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 485–492. ACM, 2006.

[7] E. Bakshy, J. Hofman, W. Mason, and D. Watts. Identifying influencers on twitter. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, 2011.

[8] S. Banerjee, K. Ramanathan, and A. Gupta. Clustering short texts using wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM, 2007.

[9] G. Barbier and H. Liu. Information Provenance in Social Media. *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 276–283, 2011.

[10] D. Carmel, H. Roitman, and N. Zwerdling. Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 139–146. ACM, 2009.

[11] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *ACM SIGMOD Record*, volume 27, pages 307–318. ACM, 1998.

[12] H.-H. Chen, M.-S. Lin, and Y.-C. Wei. Novel association measures using web search with double checking. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1009–1016. Association for Computational Linguistics, 2006.

[13] L. Chen and A. Roy. Event detection from Flickr data through wavelet-based spatial analysis. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 523–532. ACM, 2009.

[14] B. Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.

[15] B. Danushka, M. Yutaka, and I. Mitsuru. Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 757–766, 2007.

[16] L. Denoyer and P. Gallinari. The wikipedia xml corpus. *SIGIR Forum*, 40(1):64–69, 2006.

[17] J. F
"urnkranz. Exploiting structural information for text classification on the www. *Advances in Intelligent Data Analysis*, pages 487–497, 1999.

[18] E. Gabrilovich and S. Markovitch. Feature generation for text categorization using world knowledge. In *International joint conference on artificial intelligence*, volume 19, page 1048, 2005.

[19] E. Gabrilovich and S. Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1301, 2006.

[20] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12, 2007.

[21] S. Gerani, M. J. Carman, and F. Crestani. Proximity-based opinion retrieval. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 403–410, New York, NY, USA, 2010. ACM.

[22] M. Gray, B. Team, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, and S. Pinker. Quantitative Analysis of Culture Using Millions of Digitized Books. *science*, 1199644(176):331, 2011.

[23] Z. Guan, C. Wang, J. Bu, C. Chen, K. Yang, D. Cai, and X. He. Document recommendation in social tagging services. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 391–400, New York, NY, USA, 2010. ACM.

[24] J. Hammerton, M. Osborne, S. Armstrong, and W. Daelemans. Introduction to special issue on machine learning approaches to shallow parsing. *Machine Learning Research*, 2:551–558, 2002.

[25] F. M. Harper, D. Moy, and J. A. Konstan. Facts or friends?: distinguishing informational and conversational questions in social qa sites. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, pages 759–768, New York, NY, USA, 2009. ACM.

[26] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the international conference on Web search and web data mining*, pages 195–206. ACM, 2008.

[27] J. Hu, L. Fang, Y. Cao, H. Zeng, H. Li, Q. Yang, and Z. Chen. Enhancing text clustering by leveraging Wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 179–186. ACM, 2008.

[28] X. Hu, N. Sun, C. Zhang, and T.-S. Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 919–928. ACM, 2009.

[29] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou. Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 389–396. ACM, 2009.

[30] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[31] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao. Graph regularized transductive classification on heterogeneous information networks. *Machine Learning and Knowledge Discovery in Databases*, pages 570–586, 2010.

[32] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 297–304. ACM, 2004.

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

[34] Y. Lee, H.-y. Jung, W. Song, and J.-H. Lee. Mining the blogosphere for top news stories identification. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 395–402, New York, NY, USA, 2010. ACM.

[35] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 621–630, New York, NY, USA, 2010. ACM.

[36] D. Lewis and W. Croft. Term clustering of syntactic phrases. In *Proceedings of the 13th annual international ACM SIGIR confer-*

ence on Research and development in information retrieval, pages 385–404. ACM, 1989.

[37] C. Lin, B. Zhao, Q. Mei, and J. Han. Pet: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 929–938. ACM, 2010.

[38] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 691–700, New York, NY, USA, 2010. ACM.

[39] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the trec-2009 blog track. *Proceedings of TREC 2009*, 2010.

[40] D. Margineantu, W. Wong, and D. Dash. Machine learning algorithms for event detection. *Machine Learning*, 79(3):257–259, 2010.

[41] J. McLean. State of the Blogosphere, introduction, 2009.

[42] M. Mendoza, B. Poblete, and C. Castillo. Twitter Under Crisis: Can we trust what we RT? In *1st Workshop on Social Media Analytics (SOMA'10)*, 2010.

[43] S. Moturu. *Quantifying the Trustworthiness of User-Generated Social Media Content*. PhD thesis, Arizona State University, 2009.

[44] S. Osinski, J. Stefanowski, and D. Weiss. Lingo: Search results clustering algorithm based on singular value decomposition. In *Proceedings of the IIS: IIPWM'04 Conference*, page 359, 2004.

[45] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceeding of the 17th international conference on World Wide Web*, pages 91–100. ACM, 2008.

[46] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[47] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[48] B. Sigurbjornsson and R. Van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web*, pages 327–336. ACM, 2008.

[49] A. Stavrianou, P. Andritsos, and N. Nicoloyannis. Overview and semantic issues of text mining. *ACM SIGMOD Record*, 36(3):23–34, 2007.

[50] Y. Sun, J. Han, J. Gao, and Y. Yu. itopicmodel: Information network-integrated topic modeling. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 493–502. IEEE, 2009.

[51] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 797–806. ACM, 2009.

[52] J. Surowiecki. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Random House of Canada, 2004.

[53] L. Tang and H. Liu. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 817–826. ACM, 2009.

[54] L. Urena-Lopez, M. Buenaga, and J. Gomez. Integrating linguistic resources in TC through WSD. *Computers and the Humanities*, 35(2):215–230, 2001.

[55] N. Van House. Flickr and public image-sharing: distant closeness and photo exhibition. In *CHI'07 extended abstracts on Human factors in computing systems*, pages 2717–2722. ACM, 2007.

[56] J. Wang, Y. Zhou, L. Li, B. Hu, and X. Hu. Improving short text clustering performance with keyword expansion. In *The Sixth International Symposium on Neural Networks (ISNN 2009)*, pages 291–298. Springer, 2009.

[57] K. Wang, Z. Ming, X. Hu, and T. Chua. Segmentation of multi-sentence questions: towards effective question retrieval in cQA services. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 387–394. ACM, 2010.

[58] P. Wang and C. Domeniconi. Building semantic kernels for text classification using Wikipedia. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–721. ACM, 2008.

[59] X. Wang, L. Tang, H. Gao, and H. Liu. Discovering overlapping groups in social media. In *the 10th IEEE International Conference on Data Mining series (ICDM2010)*, Sydney, Australia, December 14 - 17 2010.

[60] X. Wang, C. Zhai, X. Hu, and R. Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the*

13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 784–793. ACM, 2007.

[61] D. Yin, Z. Xue, L. Hong, and B. D. Davison. A probabilistic model for personalized tag prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 959–968, New York, NY, USA, 2010. ACM.

[62] Z. Yin, R. Li, Q. Mei, and J. Han. Exploring social tagging graph for web object classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 957–966, New York, NY, USA, 2009. ACM.

[63] J. Yuan, Z. Zha, Z. Zhao, X. Zhou, and T. Chua. Utilizing related samples to learn complex queries in interactive concept-based video search. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 66–73. ACM, 2010.

[64] R. Zafarani and H. Liu. Connecting Corresponding Identities across Communities. In *Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM09)*, 2009.

[65] T. Zesch, C. Muller, and I. Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 1646–1652. Citeseer, 2008.

[66] Z. Zha, X. Hua, T. Mei, J. Wang, G. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[67] Q. Zhao, P. Mitra, and B. Chen. Temporal and information flow based event detection from social text streams. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, pages 1501–1506. AAAI Press, 2007.

[68] Y. Zhou, H. Cheng, and J. Yu. Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, 2(1):718–729, 2009.