Research Journal of Applied Sciences, Engineering and Technology 6(21): 4070-4076, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: January 29, 2013 Accepted: March 21, 2013 Published: November 20, 2013

A New Method for Extracting Key Terms from Micro-Blogs Messages Using Wikipedia

Ahmad Ali Al-Zubi
King Saud University, P.O. Box (28095), Riyadh 11437, Saudi Arabia

Abstract: This study describes how to extract key terms of the micro-blogs messages, using information obtained by analysing the structure and content of online encyclopaedia Wikipedia. The algorithm used for this target is based on the calculation of "*keyphraseness*" for each term, i.e., assess the probability that it may be chosen as a key term in the text. During assessment, the developed algorithm has shown satisfactory results in terms of this task, significantly outpacing other existing algorithms. As a demonstration of the possible application of the developed algorithm it has been implemented in a system prototype of contextual advertisement. And some options have been also formulated using the information obtained by analysing Twitter messages, for various support services.

Keywords: Blogs, messages, extracting algorithms, key term extraction, keyphraseness, micro blogging

INTRODUCTION

Key Terms extraction is vital for Knowledge Management Systems, Information Retrieval Systems and Digital Libraries as well as for general browsing of the web. Key terms are often the basis of document processing methods such as clustering and retrieval since processing all the words in the document can be slow. Common models for automating the process of key terms extraction are usually done by using several statistics-based methods such as Bayesian, K-Nearest Neighbor and Expectation-Maximization. These models are limited by word-related features that can be used since adding more features will make the models more complex and difficult to comprehend (Arnulfo *et al*., 2012).

To date, one of the most important and visible areas of Web 2.0, which's key principle is the user involvement in the study of sites are online diaries, or Web logs, called "blogs". Conceptual development of blogs is due to their broad socialization, are micro blogs, which have certain characteristics: a limited message length, high frequency of publication, various topics, different ways of delivering messages, etc.

The first and best-known micro blogging service "Twitter" was launched in October 2006 by the company "Obvious" from San Francisco. To date, the constantly growing audience of this service is reaching tens of millions of people. Obviously, the automated selection of the most significant terms of the flow of messages generated by the community of Twitter, has a practical importance for determining the interests of different groups of users, as well as to build an individual profile of each of them. Over the recent years a number of studys were published on key terms extraction using different approaches (Mihalcea and Csomai, 2007a). Al-Zubi (2010) and Mihalcea and Csomai (2007b) Give a detailed overview of many existing methods for key terms extraction using "Wikipedia".

However, it should be noted that the classical statistical methods for the extraction of key terms, based on the analysis of document collections are ineffective in this case (Al-Zubi, 2010). This is due to the extremely small length of messages (up to 140 characters), their wide range of themes and the lack of logical connection between them, as well as an abundance of low use of abbreviations, acronyms and elements of specific micro-syntax.

In this research study, to solve this problem we determine the relative importance of terms in the analyzed context by the data on the frequency of their use as keys in the online encyclopedia "Wikipedia". Our algorithm is based on the calculation of "*keyphraseness*" of each term, i.e. assess the probability that it may be chosen as a key in the text. Further, a number of heuristics is applied to the analyzed set of terms, which's result is a list of terms found to be keys.

MATERIALS AND METHODS

Review of twitter: Modern internet researchers believe that, the emergence and subsequent development of the idea of a micro blogging service is a quite reasonable result of the integration of the concept of social networking with online diaries - blogs (Martin, 2008).

According to the definition given by Walker (Herman *et al*., 2005), known as "blogs are frequently updated websites", that consist of a variety of records containing information that, are placed in reverse chronological order.

Table 1: Concepts interaction modes

	Channel	Element
Channel	Follow	Retweets
Element	@-Link	Reply

Characteristic features of blogging can be described using three key principles (Böhringer, 2009):

- The content of blogs is a short message
- Messages have a common authorship and are controlled by the author
- Possible aggregation of multiple streams of messages from differentauthors for easy reading

These principles are also applicable to micro-blogging (Karger and Quan, 2005). However, while both the publishing and aggregation of posts for blogging are considered tasks for different software products, services of micro blogging provide all these features at once.

In addition, micro-blogging takes into account users need in a faster mode of communication than usual blogging. By encouraging more short messages, it reduces the time and mental work needed to create the content. It is also one of the main distinguishing features of blogging. Another difference lies in the frequency of updates. Author usually updates his blog, on average, once every few days, while the author micro blog can update it several times a day.

The main functions of the most popular to date, micro blogging service "Twitter" has a very simple model. Users can send short messages or tweets, no longer than 140 characters. Messages are displayed as a stream on the user page. In terms of social networks, Twitter allows users to follow any number of other users, called friends. Twitter network of contacts is asymmetric, i.e., if a user follows another; the latter is not obliged to follow him. Members following another user are called his followers.

Users have the opportunity to indicate whether they wish to have their tweets publicly available (they are in reverse chronological order on the main page of the service and on the user's page, called micro blog) or privately (only followers of the user can see his posts). By default, all messages are available to any user.

To facilitate understanding the internal structure of Twitter, it is useful to introduce two concepts: the element as a separate message and channel as a flow of elements, mostly belonging to one user (Böhringer, 2009). Concepts interaction modes are presented in Table 1.

Below is a brief description of each interaction mode:

- **Follow:** One channel has a different channel in its network and reading its updates
- **@-link:** The element's text may refer to another channel by the construction @ <channel's_name>
- **Retweet:** Users taking elements from other people's channels and putting them in their channels with the addition of @-references to the source and, in some cases, their own comments
- **Reply:** One element is a direct response to the previous one

Generally describing the features of Twitter, we can note that a necessary condition for the successful adoption of new technology by users (or a new way to use existing tools) is a positive attitude towards its potential. , Gartner added micro blogging to its "hype cycle" in Gartner Highlights (2008) 27 Technologies in the 2008 Hype Cycle for Emerging Technologies, 2008, predicting a sharp rise in the popularity of this phenomenon. According to Gartner, leading companies exploring the potential of micro blogging to improve other social information tools and channels. It all say that micro blogging is one of the most promising and fastest growing segments of the Internet.

However, micro blogs are still a relatively new phenomenon of online social networks and at this stage not been investigated.

Posts features: In addition to the presence of "retweets" and @-links in the "tweets" there can also be present another items of a specific micro-syntax, whose purpose is to provide frequently used concepts in abbreviated vernacular form as well as expand the range of tools to enhance the informative messages in a limited size.

The main part of micro-syntax is slash tags, each of which consists of the symbol "/" and the index, which defines the meaning of slash tag. These elements are used for different purposes. For example, «/ via»-a reference to the author for "retweet»; «/ by»-a reference to the author's original message, if it is the result of a chain of "retweets»; «/ cc»-to indicate those subscribers of micro blogs, to which messages are primarily addressed, etc.

Using a fully slash tags is an initiative of the users, so there are no specific rules for their use. The above description reflects the most popular to date ways to use slash tags. However, there are other recommendations that are also noteworthy, since each user uses micro-syntax items at his discretion. For example, it is possible to group all used slash tags in the message without symbol "/" in one group.

Because in Twitter it is not simple and convenient to group "tweets" on the same subject of different users, the users' community has come to its own solution: Use hash tag. They are similar to other examples of the use of tags (for example, to annotate records in the ordinary blogs) and allow you to add "tweets" in any category.

Hash tags begin with a "#" followed by any combination of allowed non-space characters in Twitter; most often it is a word or phrase in which the first letter of each word is an upper case. They can occur in any part of the "tweets", often users simply add the "#" in front of any word. Another way is to add a

popular hash tag, such as «# haiku». When you add a hash tag to the message it will be displayed when searching in a messages stream of Twitter by the hash tag.

Existing approaches for key terms extraction: One of the tasks of extracting information from text is a selection of key terms, with a certain degree of confidence reflecting the thematic focus of the document. Automatic extraction of key terms can be defined as the automatic selection of important thematic terms in the document. It is one of the subtasks of a more general problem-the automatic generation of key terms, for which the selected key terms do not have to be present in this document (Turney, 1999). For the last years there have been many approaches that allow the analysis of group of documents of various sizes to extract key terms, consisting of one, two or more words.

The most important step in extracting key terms is calculating their weights in the targeted document that allows you to evaluate their importance relative to each other in this context. To solve this problem, there are many approaches that are divided into two groups: requiring training and not requiring training. Under “training” we mean the need for pre-processing the source text corpus in order to extract information about the frequency of occurrence of terms in the whole body. In other words, to determine the significance of the term in this document, you must first analyze the entire collection of documents to which it belongs. An alternative approach is the use of linguistic ontologies, which are more or less approximate models of the existing set of words of a given language. On the basis of both approaches some systems have been developed for the automatic key terms extraction, but some works are still ongoing in this direction to improve the accuracy and completeness of the results, as well as to use the methods of extracting information from text to address new problems (Turdakov, 2010; Turdakov and Kuznetsov, 2010; Dmitry *et al*., 2010; Lizorkin *et al*., 2008; Grineva *et al*., 2009a, b; Turdakov and Lizorkin, 2009).

The most common schemes for calculating the weights of terms are the TF-IDF and its various options, as well as some others (ATC, Okapi, LTU). However, a common feature of these schemes is that they require information from the entire collection of documents. In other words, if a method based on TF-IDF, used to create a presentation about the document, the arrival of a new document in to the collection requires a recalculation of the weights of terms in all documents. Therefore, any application based on the values of the weights of terms in the document will also be affected. This largely precludes the use of key terms extracting methods that require learning, in systems where dynamic data streams must be processed in real time, for example, messages of micro blogs (Grineva *et al*., 2009).

Several approaches have been proposed to solve this problem, such as an algorithm TF-ICF (Reed *et al*., 2006). As a development of this idea Mihalcea and Csomai (2007a) proposed to use Wikipedia as a training thesaurus (Mihalcea and Csomai, 2007b). They used to calculate the information contained in the annotated encyclopedic entries with hand-selected key terms. To estimate the probability that the term will be chosen as a key in the new document, this formula can be used:

$$P(\text{Key Term}|W) \approx \frac{\text{Number } (D_{key})}{\text{Number } (D_W)} \quad (1)$$

where,

W = A term
D_{key} = A document in which the term has been chosen as a key
D_W = A document in which the term has appeared at least once

This assessment has been named by the authors “keyphraseness”. It can be interpreted as follows: "The more often the term was chosen among its total number of occurrences, the more likely it will be selected as such again."

Keyphraseness can take values from 0 to 1. The higher it is, the higher the significance of the term in the analyzed context. For example, for the term «of course» on Wikipedia, it can only be found in one article, so it is rarely selected as a key, there for the value of its keyphraseness will be close to 0. On the contrary, the term «Microsoft» in the text of any study will usually be selected as a key, so the value of its keyphraseness will be very close to 1.

This approach is quite accurate, since all studies on Wikipedia manually annotated key terms and therefore the proposed assessment of their actual keyphraseness is only the result of processing people's opinions.

However, this estimate may be unreliable in cases when values used for calculation are too small. To solve this problem, the authors recommend to consider only those terms that appear on Wikipedia for at least 5 times.

As a conclusion of the revision of methods for key terms extraction, to calculate the weight of the terms in this study we used this equation:

$$W_i = TF_i * K_i \quad (2)$$

where,

i = Ordinal number of the term
TF_i = The frequency of the term in the message
K_i = Keyphraseness of the term in Wikipedia

TF means Term Frequency. Value of this component of the formula is the ratio of the number of some term occurrences to the total number of terms the message. Thus, the importance of term t_i is evaluated within a single message.

RESULTS AND DISCUSSION

Extracting information from Wikipedia: One of the most important stages of system development was the processing of XML-dump of English Wikipedia studies as of July 2009 The purpose of the analysis was the calculation of keyphraseness for all of the terms of Wikipedia using Eq. (1).

It should be noted that for one concept in Wikipedia dictionary could be found several synonyms. For example, the term «IBM» has several synonyms: «International Business Machines», «Big Blue», etc. Since the developed system is not designed to permit stage of lexical ambiguity of terms, it was unacceptable for the synonyms to have different values of keyphraseness. Therefore, it was assumed that the keyphraseness of all synonyms of the same of the same concept is equal, based on the overall statistics for all of them.

In addition, as recommended by the authors of key phraseness calculation method (Mihalcea and Csomai, 2007a), were excluded terms that were found in less than five studies. If you skip this step, the resulting value is often unreliable and cannot properly assess the relative importance of the term in context. As a result of this step the DB contains 8 435 667 terms with calculated keyphraseness.

Key terms extraction: The general architecture of the developed system is shown in Fig. 1.

To get information about user messages from Twitter server is been used Perl-modul Net : Twitter. Because of interaction with the Twitter API is getting a number of recent posts user's account, known as timeline.

To solve this problem we have chosen the method of statuses/ friends_timeline, which returns the friends message of user. For testing, we have created accounts, to update the status of only one friend. In addition, at the same time in the same account was not published any messages. Thus, the result of calling this method provides only the necessary number of messages per Twitter user, what is required as input data.

During pre-processing of text, the contents of the received message from Twitter server converted to input format of the algorithm for extracting key terms.

Besides the standard operations for this stage, can also be performed the removal of slash tags and @-links, as well as function words «RT» and «RETWEET», that is, those elements of the specific syntax of Twitter, which do not carry meaning and are considered as a stop-words in terms of processing

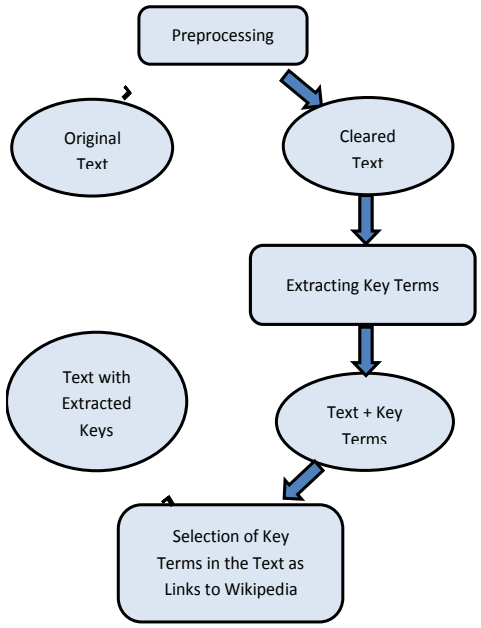

Fig. 1: General system architecture

natural language. At this stage also is performed the extraction of hash tags, which are subsequently processed along with the other words.

However, clearly, that the list of candidates for key terms is not set up only of the original words sequence. Based on the short length of messages, it is possible to make the algorithm as surplus to improve the completeness of the results that is not to miss any possible key term, how many words it may be composed of. For this purpose, a search for all possible ***N***-grams is made. The number of ***N***-grams Q_k, which can be obtained from k words (including N-grams of a single word), so that is:

$$Q_k = \sum_{i=1}^{k} i \quad (3)$$

All obtained terms at this stage are added to the array of possible key terms.

The final stage of preprocessing is stop listing, i.e., remove those words that do not carry big semantic value from the resulting array. It is important to note that stop listing is satisfied only after removal of ***N***-grams. Thus, the stop-words are included in compound terms, but removed from the list of candidates. At this stage the stop-list of SMART system can be used (Salton, 1971).

At the stage of calculating the weights of the candidate terms, the value of keyphraseness for each of them is inquired from the database. The second necessary indicator for the calculations is occurrence frequency of the term TF. The weight for each term found in the database is calculated by Eq. (2).

The general principle of key terms extraction is to analyze a given number of messages and determine the threshold weight for each of them. Those terms of

weight greater than or equal to threshold values, are considered as a key.

Initially, the threshold is considered to be the arithmetic average of the weight value of all candidate terms. All subsequent operations are designed to refine it and improve the results of the algorithm as a whole.

The next step is to process the array of hash tags obtained during preprocessing. It is assumed that with their help, the user explicitly specifies the terms that define the subject of the message. It is therefore logical to assume that the threshold value for the entire message must not be higher than the minimum weight among its hash tags. Based on this assumption, the weight of each of the hash tags (if it has been found in the database) is compared with the current threshold value and it is reduced if it is greater.

If, after processing hash tags threshold value remains equal to 0 (when they were not listed or if none of them was not found in the database) or greater than the average value that was found, it is assumed to be equal to the average. This situation often occurs in practice, because users rarely clearly indicate thematic terms. Otherwise, this approach significantly improves the results of this study.

It should be noted that the messages are processed in reverse order of arrival from the server that is in direct chronological. This approach is logical and takes into account specificity of blogging service as a whole: the user can write a message on any topic and then return to it again. However, the second message, in addition to those key terms that have been selected in the first one, there may be other, more informative terms by which the threshold for the second message will be higher and the key terms from the first message will be highlighted. To avoid this, the system has a separate array containing all the previously extracted key terms. Then while processing the next message the terms of this array will be extracted and certainly reduces the weight threshold for a given message.

In this context, it is important that if you directly select key terms the candidates are processed in ascending order of their weights. Thus, if the weight of any of the candidates below the threshold, but the key is chosen because of its presence in the array of previously extracted terms, the threshold value becomes equal to its weight and all of the following terms are automatically placed in the list of key terms.

The result of the algorithm is a list sorted in descending order of weights of key terms.

Interaction with the Amazon API: To demonstrate the possible application of the developed algorithm, it was implemented to obtain product descriptions from Amazon store server, relevant to found key terms.

To interact with the Amazon REST API we used Perl-module Net:: Amazon, which offers convenient access to most functions of the software interface. At the same time it searches for all products of the online store, in which's title or description that meets the search term. For the required number of the most suitable products are derived the name, price, year of publication and the image. Product Name is a link to its page on Amazon website.

When connecting to the server a developer private key is used, which is provided when registering to the service from Amazon and allows conducting transition statistics on products page from third-party sites? If, after following a link, the user gets the product, then according to the affiliate Amazon program, the site owner can get a reward equals to part of the product cost. Such a scheme can be implemented not only for Amazon, but also for any other online store that has an affiliate program.

The experimental results: The output of the system is an HTML-page, divided into several blocks, each of which corresponds to one message. The block displays the text of the original message with its author name, then-the text after preprocessing and finally, the same text after processing. In message's text at the output of the system the found key terms are links to relevant Wikipedia studies.

For all found key terms a table is being constructed, each row of which contains a term, its weight and found relevant items from the online store. And also the mean and the threshold values of the weight. The last part of the output is a list of terms that were found in the database, but were not assigned as a key.

The effectiveness of key terms extracting algorithms is usually evaluated by comparing the results of their work with the manual key terms extraction. Performance is measured based on the number of correspondences between sentences extracted by the algorithm and manually (Turney, 1999).

To test the system a few test accounts were created, each of which was subscribed to update the status of various-well known in the IT-community users of Twitter. Semtweettest2 has been selected as the main account for testing, which was "signed" to update the blog of Tim O'Reilly (timoreilly), a publisher and public figure who has more than 1.4 million subscribers. Posts in this blog are of an extremely diverse subjects, they often use different named entities (names of individuals, companies and events, locations names), which are of real interest at this time. In addition, the author of the blog fully uses the advantage of Twitter micro-syntax. All of what we have mentioned above gives the base to assume that the results of the developed system for the posts of timoreilly blog can reliably assess the efficiency of the algorithm.

Table 2: The results of the system testing

Method	Accuracy %	Completeness %	F-measure %
Alchemy API	22.6	44.8	27.3
Developed system	43.2	69.9	52.7

Alchemy API system was selected to compare the results of the algorithm with the existing analogs (Alchemy API), which provides access to its demo features in the online mode. As an initial data, the system uses a text document and returns the same document with the extracted key terms.

Each of the selected test messages was analyzed using the developed system and a demo version of Alchemy API. Since there is no need to calculate the accuracy, completeness and F-measure for each of them, the whole array of posts was adopted as a single document from which the key terms were extracted.

Of the selected 50 messages, a total of 180 key terms has been extracted manually, 28 of which are parts of other, more long-term. The maximum length of manually extracted terms is 3 words. While the maximum length of the term extracted by the system is 6 words.

The test results are shown in Table 2.

Based on the test results can be concluded that the developed system is functioning effectively in the conditions of the given task. In addition, in terms of quality of the results, it is superior to the selected Alchemy API system for comparison.

One possible reason for reducing the quality of the results is the large number of named entities in the processed messages text, most of them refer to the people, events or companies that have become only recently popular. Another reason is the frequent use of acronyms, many of which are not common.

CONCLUSION

As a result of this study, an algorithm was developed for extracting key terms of the minimum micro-structured text messages. The conducted experiments showed that the algorithm studies correctly and efficiently. As an example of possible practical applications of the results within the framework of the developed system, a system prototype was implemented of contextual advertising of products from the an online store.

As a part of the future study I am planning to create service, based on OAuth authentication mechanism, which is used in most modern applications for Twitter. User also provides access to the data of his own account, including the text of all messages. One of the possible prospects of using the algorithm is the ability for the user to specify several key terms in order to see only the messages that contain them. Filtering is possible not only for extracted terms through Wikipedia, but also by a simple search through text messages after preprocessing. Such a service would be in demand (Zhao and Rosson, 2009) and would allow to attract the audience to display contextual advertising.

REFERENCES

Al-Zubi, A.A., 2010. A hybrid method for extracting key terms of text documents. Int. J. Video Image Proces. Network Security, 10(02): 8-13.

Alchemy API–Demo. Retrieved from: -http://www.alchemyapi.com/api/demo.html.

Arnulfo, A., M.D. Liu and R. Setiono, 2012. Keyword extraction using back propagation neural networks and rule extraction. WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, Brisbane, Australia.

Böhringer, M.R., 2009. Social syndication: A conceptual view on micro blogging. Sprouts: Working Papers on Information Systems, 9(31).

Dmitry, L., P. Velikhov, M. Grinev and D. Turdakov, 2010. Accuracy estimate and optimization techniques for sim rank computation. Int. J. Very Large Data Bases Arch., 19(1).

Gartner Highlights, 2008. 27 Technologies in the Hype Cycle for Emerging Technologies. Retrieved from: http://www.gartner.com/it/page.jsp?id=739613.

Grineva, M., M. Grinev, A. Boldakov, L. Novak, A. Syssoev, D. Lizorkin, 2009a. Sifting micro-blogging stream for events of user interest. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval.

Grineva, M., M. Grinev and D. Lizorkin, 2009b. Extracting Key Terms From Noisy and Multitheme Documents. -WWW2009: 18th International World Wide Web Conference.

Grineva,M.,M.Grinev and D. Lizorkin, 2009c. Effective extraction of thematically grouped key terms from text. Proceeding of the AAAI 2009 Spring Symposium on Social Semantic Web, pp: 39-44.

Herman, D., J. Manfred and R. Marie-Laure, 2005. The Routledge Encyclopedia of Narrative Theory. London, Routledge.

Karger, D.R. and D. Quan, 2005. What would it mean to blog on the semantic web? Web Semantics: Science, Services and Agents on the World Wide Web, Selected Papers from the International Semantic Web Conference, Hiroshima, Japan, 07-11 November 2004, 3(2-3): 147-157.

Lizorkin, D., P. Velikhov, M. Grinev and D. Turdakov, 2008. Accuracy estimate and optimization techniques for simrank computation-Proceedings of the VLDB Endowment, 1(1).

Martin, E., 2008. Microblogging-more than fun?-Proceedings of IADIS Mobile Learning Conference 2008, Inmaculada Arnedillo Sánchez and Pedro Isaías ed., Portugal, pp: 155-159.

Mihalcea, R. and Csomai, A. 2007a. Wikify!: linking documents to encyclopedic knowledge. Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, New York, USA, ACM, pp: 233-242.

Mihalcea, R. and A. Csomai, 2007b. Wikify!: Linking documents to encyclopedic knowledge. Proceedings of the 16th ACM Conference on Information and Knowledge Management, ACM Press, New York, USA., pp: 233-242.

Reed, J.W., Y. Jiao, T.E. Potok, B.A. Klump, M.T. Elmore and A.R. Hurson, 2006. TF-ICF: A new term weighting scheme for clustering dynamic data streams. Proc. Machine Learning and Applications, ICMLA '06, pp: 258-263.

Salton, G., 1971. The SMART Retrieval System-Experiments in Automatic Document Processing. Prentice-Hall Inc., Englewood Cliffs, NJ.

Turdakov, D. and D. Lizorkin, 2009. HMM expanded to multiple interleaved chains as a model for word sense disambiguation. PACLIC 2009: The 23rd Pacific Asia Conference on Language, Information and Computations, pp: 549-559.

Turdakov, D. and S. Kuznetsov, 2010. Automatic word sense disambiguation based on document networks. Program. Comput. Softw., 36(1): 11-18.

Turdakov, D., 2010. Word sense disambiguation methods. Program. Comput. Softw., 36(6): 309-326.

Turney, P., 1999. Learning to extract key phrases from text. Technical Report, National Research Council, Institute for Informational Technology.

Zhao, D. and M. Rosson, 2009. How and why people twitter: The role that micro-blogging plays in informal communication at work. Proceedings of the ACM 2009 International Conference on Supporting Group Work.