

A Spatio-temporal Framework for Related Topic Search in Micro-Blogging

Shuangyong Song, Qiudan Li, and Nan Zheng

Laboratory of Complex Systems and Intelligence Science

Institute of Automation, Chinese Academy of Sciences

95 Zhongguancun East Road, Beijing, China 100190

{shuangyong.song,qiudan.li,nan.zheng}@ia.ac.cn

Abstract. With the rapid development of Web 2.0, micro-blogging such as twitter is increasingly becoming an important source of up-to-date topics about what is happening in the world. By analyzing topic trends sequences and identifying relations among topics we have opportunities to gain insights into topic associations and thereby provide better services for micro-bloggers. This paper proposes a novel framework that mines the associations among topic trends in twitter by considering both temporal and location information. The framework consists of the extraction of topics' spatio-temporal information and the calculation of the similarity among topics. The experimental results show that our method can find the related topics effectively and accurately.

1 Introduction

Micro-blogging is a Web2.0 technology and a new form of blogging. Compared to regular blogging, micro-blogging realizes an even faster mode of communication [10]. It allows users to publish short message updates in different channels, including the Web, SMS, e-mail or instant messaging clients. Every update is usually limited to 140-200 characters, sometimes images and audios are added to enrich its contents. Users in micro-blogging are dubbed micro-bloggers, and the short message updates published by the users are dubbed micro-blogs. Unlike other social network services, in micro-blogging, a user A is allowed to "follow" other users without seeking any permission, and in real time the updates of the followed users will be sent to A automatically. If A is following B, B is call A's friend, and A is called B's follower. Thus friendship can either be reciprocated or one-way.

During the last few years, micro-blogging has become one of the most popular Web 2.0 services, and it is still growing rapidly. Take Twitter¹, a popular micro-blogging website, as an example, Nielsen.com reports that the total number of users in Twitter has increased from 530,000 in Sep. 2007 to 2,360,000 in Sep. 2009 [1]. A great deal of media attention has been focused on Twitter. On April 17, 2007, the day when Oprah Winfrey joined Twitter by sending a tweet from her Friday TV show, shared of US based visits to the Twitter site increased by 24% and some 1.2 million new users signed up for Twitter on that day alone [9]. In May, 2007, the White House

¹ <http://www.twitter.com>

began posting short messages on Twitter [11]. Figure 1 shows a snapshot of Twitter's user interface². Like the regular blogging service interfaces, the user's registration information and her posts are shown on her homepage, where the latest tweet she published is with bigger font than the previous ones. In addition, some other application modules are shown in this interface: number of friends, followers, related lists and publisher tweets; the users who she is following; whether she is followed by the user who is looking at her homepage now. Those differences between micro-blogging and regular blogging is just because of the novel interactive mode between users in micro-blogging, 'follow'.

Changing the official question from "What are You Doing?" to "What's Happening", Twitter has becoming an important source of up-to-date topics about what is

I am following her.

The latest tweet

Previous tweets

User information

Number of friends, followers, related lists and published tweets.

Following (Friends)

Fig. 1. An example of Twitter's user interface

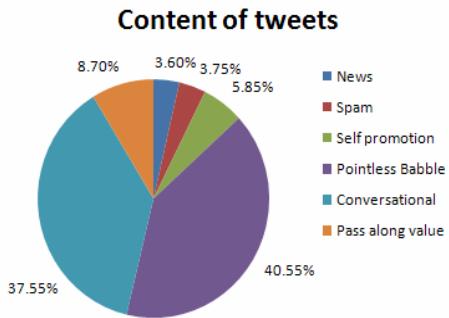


Fig. 2. Content areas of Twitter

Country	Percent of Site Traffic
United States	35.4%
India	8.0%
Germany	6.8%
United Kingdom	6.1%
Japan	5.9%
Brazil	2.9%
Canada	2.5%
Indonesia	2.0%
Australia	1.9%
Netherlands	1.7%

Fig. 3. The geographical distribution of users in Twitter

² <http://www.twitter.com/songshuangyong>

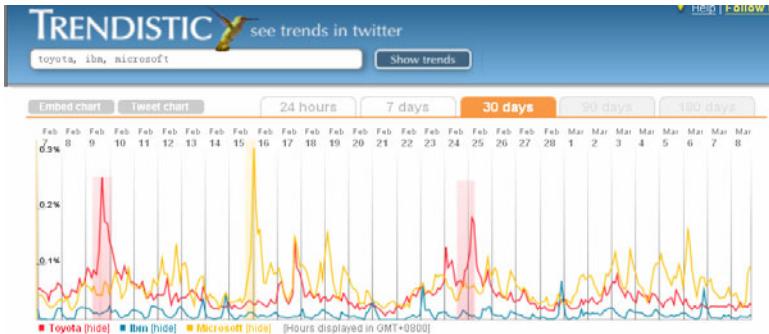


Fig. 4. The trends of ‘Toyota’, ‘IBM’ and ‘Microsoft’ in Twitter

happening in the world. We analyze the content of tweets in Twitter with statistical data given in [19], which is shown in figure 2. The results are interesting, Conversational (such as questions or polls.) and Pointless Babble (the tweets like “I am eating a sandwich now.”) account for more than 3/4 of all contents. This result proves that Twitter is mainly used to note what was happening around the users and to communicate with other people, and most of the topics on Twitter are about the users’ daily life, such as food, music and electronic products. The geographical distribution of users in Twitter is given in figure 3, which is downloaded from Alexa.com³. From figure 3, we can see that more than one third of users in Twitter are from the United States, where the Twitter launched. Twitter is most popular in US, Europe and Asia (mainly India and Japan).

Trends of three topics, ‘Toyota’, ‘IBM’ and ‘Microsoft’⁴, are given in figure 4 to describe the temporal information of topics in Twitter, which we will use to calculate the similarity between topics. We can see that ‘Toyota’ and ‘Microsoft’ always got more attention than ‘IBM’ from Feb. 7th, 2010 to Mar. 8th 2010. On Feb. 16th, 2010, the tweets about ‘Microsoft’ account for 0.3% of all the tweets. This is because that Microsoft was publicly previewing Windows Phone 7 for the first time on that day.

As shown, real-time topics have significant temporal and location aspects, which are not taken into consideration by most micro-blog search engines. Through the spatio-temporal information, we can easily analyze the periods of time when topics have been concerned and the regional distributions of the micro-bloggers who have posted micro-blogs on those topics. Since every topic can be defined and described with their temporal and spatial information [6], the relationship among those topics can also be detected by calculating the similarity between any two of them through their spatio-temporal information.

In this paper, we propose a novel framework to identify correlations among topics with their temporal and location components. We present a similarity-based searching and pattern matching algorithm that identifies spatio-temporal series data with similar temporal dynamics and location statistics (the location information of the micro-bloggers who have posted micro-blogs) in a specific period of time. By analyzing topic

³ <http://www.alexa.com/siteinfo/twitter.com#demographics>

⁴ http://trendistic.com/toyota/ibm/microsoft/_30-days

trends sequences and identifying relations among topics we have opportunities to gain insights into topic associations and thereby provide better services for micro-bloggers.

The rest of this paper is organized as follows: In section 2, we provide a brief review of the related work. The introduction of our method was proposed in section 3. In section 4, some analysis of our experimental results is given. Finally we make a conclusion and discuss our plans for future work in section 5.

2 Related Work

2.1 Related Topic Detection

Related topic detection attracts a lot of attention from researchers with the growth of online search, and the related technique has been explored in previous work such as query expansion and keyword suggestion. In [3], a new framework was proposed for performing better semantic related search suggestions with complex semantic relatedness, using the real time Wikipedia-based social network structures. A frequently used approach, co-occurrence, was tested on a dataset collected from eBay website (www.ebay.com) in [14] to recommend the sellers relevant and informative terms for title expansion. Besides, three particular features, including concept term, description relevance and chance-to-be viewed, was taken into account in the application scenario. In [16], Ribeiro-Neto et al. proposed several strategies and a term expansion method to seek the relationship between the advertisements and the web pages. Recently, utilizing search engines to help find the ontology alignment is another type of method to weight approximate topic matches [20]. For example, Gligorov et al. [20] proposed a method based on the search results from Google to find the alignment between topics.

Another application of related topic detection is query expansion, which is applied when users expect additional keywords to achieve relevant documents and filter out the irrelevant ones. In [17], Buckley et al. extracted terms from known relevant documents or the top retrieved documents to add some terms to the original query. In [18], Xu and Croft proposed local context analysis, which combined the advantages of global and local expansion techniques. In [7], Mitra et al. retrieved an initial set of possibly relevant documents, and discovered correlated features to expand the query. In [8], Qiu and Frei expanded queries by adding those terms that are most similar to the concept of the query, rather than selecting terms that were similar to the query terms. In [5], Cui et al. tried to find co-occurrences with the seed term in query log.

2.2 Spatio-temporal Model

Mining topics from web-based text data and analyzing their spatio-temporal patterns have applications in multiple domains. In [12], Lu et al. proposed a novel spatio-temporal model for collaborative filtering applications, which was based on low-rank matrix factorization that used a spatio-temporal filtering approach to estimate user and item factors. In [6], Li et al. proposed a probabilistic model to detect retrospective news events by explaining the generation of “four Ws” - who (persons), when (time), where (locations) and what (keywords), from each news article. However, their work considered time and location as independent variables, and aimed at discovering the

reoccurring peaks of events rather than extracting the spatiotemporal patterns of themes. Model construction for mining spatiotemporal theme patterns from weblog data was also investigated in [13]. The authors used a probabilistic approach to model the subtopic theme and spatiotemporal theme patterns simultaneously. In [15], Syeda-Mahmood et al. tried to automatically characterize the spatio-temporal patterns in cardiac echo videos for disease discrimination using prior knowledge of the region layout.

Different from the previous work, we apply the spatio-temporal model to detect the related topics in micro-blogging. We describe topics with their temporal and spatial information, and detect the relationship among them by calculating the similarity between any two of them through their spatio-temporal information.

3 Our Approach

Figure 5 shows the system architecture of our correlated topics search framework. The ‘Tweets’ means posts broadcasted by users in Twitter about small things happening in their daily life, like what they are thinking and experiencing, and the “User information” is filled in by users with their names, regions or some other private information. Those two aspects are used to generate the location information of topics by statistic of region distributions of users. The ‘trends data’ means the Twitter’s daily trending topics about what is happening in the world, which is used to generate the temporal information. We detect related topics with the topic issued by a user by comparing their spatio-temporal information. The most important parts in this framework are the extraction and representation of topic information and similarity calculation among topics.

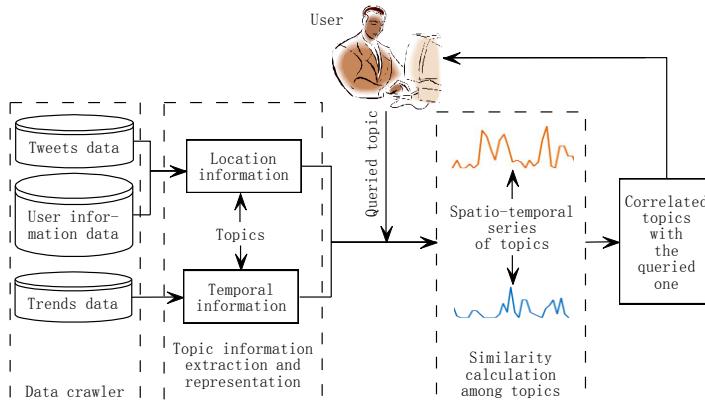


Fig. 5. Spatio-temporal framework for related topic search

3.1 Data Crawler

We use the Twitter trends API⁵ to download the statistics of the everyday trending topics on Twitter, where the topics are given in the form of popular queries. From Jan.

⁵ <http://apiwiki.twitter.com/Twitter-Search-API-Method:-trends-daily>

1st to Dec. 31st, 2009, there were 171,735 queries which contain 17,619 topics. The most popular topic ‘Red’ appeared 2,414 times, while lots of topics just had once. The distribution of the frequencies per topic is shown in figure 6. We use those data to generate the everyday frequency of the chosen topics, which stands for the temporal dynamics. Then we download the tweets dataset and the users’ information dataset published by Munmun De Choudhury⁶ to generate location information of each topic. We extract all location names and statistic their frequencies.

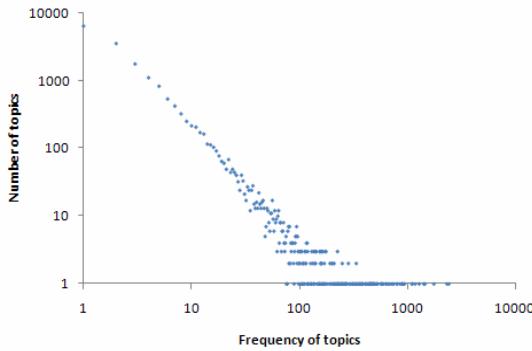


Fig. 6. Distribution of the frequencies per topic

3.2 Topic Information Extraction and Representation

The spatio-temporal topic has temporal and location aspects. The temporal information is referring to the frequency behaviors of a topic at predefined time periods, while the location information is referring to the regional distribution of the micro-bloggers who talked about a topic.

Inspired by the idea in [2], we define a topic as the following state series:

$$\text{Topic}_k = [t_{k1}, t_{k2}, \dots, t_{ki}, \dots, t_{kT}, s_{k(T+1)}, s_{k(T+2)}, \dots, s_{k(T+j)}, \dots, s_{k(T+S)}]. \quad (1)$$

where t_{ki} represents the frequency state of topic_k at timestamp i, and $s_{k(T+j)}$ represents the frequency state of the micro-bloggers, who have posted tweets on topic_k, in the region j. T in the definition means the total number of days in our chosen period of time, and S means the number of regions where tweets on topic_k have been posted in the same period of time.

In figure 7, the temporal dynamics and regional distributions of topic ‘Microsoft’ are transformed into state series, of which each dimension is an integer between 0 and 5. From the curve, we can see that temporal frequency of ‘Microsoft’ is always high in the first two months in 2009, and the Pacific Time Zone is the region where the largest number of tweets about ‘Microsoft’ had been posted in this period of time.

Representing topics as state series in equation 1 and calculating the similarity among them, we omit taking into account a topic’s possible co-existence with another

⁶ <http://www.public.asu.edu/~mdechoud/>

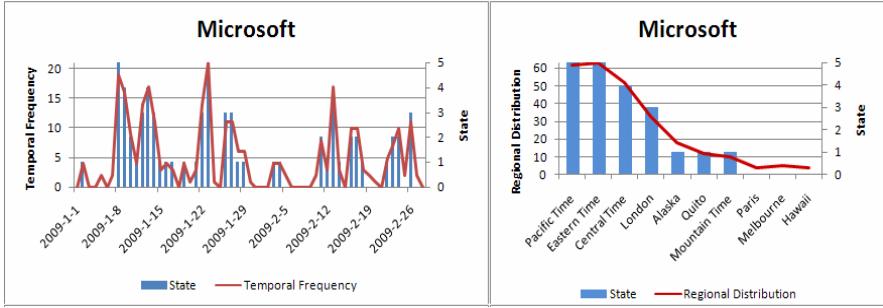


Fig. 7. State series of ‘Microsoft’

topic in the same tweet [2, 4]. Considering the spatio-temporal similarity of those topics enables us to find implicit association among them although they may not appear in the same document.

3.3 Similarity Calculation among Topics

We design a similarity measurement, TS α ED to calculate the similarity between two topics by comparing both their temporal dynamics and their regional distributions. Correlated topics have similar spatio-temporal characteristics, and to some extent, our method takes into account topics’ possible co-existence with each other in the same micro-blog.

The Euclidean distance metric is adopted to compute the similarity between two topics A and B, which have been represented as state series. The final formula is given as follows:

$$TS\alpha ED(A, B) = \alpha \sqrt{\sum_{i=1}^T (A_i - B_i)^2} + (1 - \alpha) \sqrt{\sum_{s=1}^S (A_{T+s} - B_{T+s})^2} \quad (2)$$

where A_t means the frequency of topic A in the t^{th} day, A_{T+s} means the numbers of users in s^{th} region who have posted tweets about topic A. The parameter α is used to adjust the significance of sequence similarity and location similarity, and finally α is chosen to be 0.59 after the cyclic iterative method.

4 Experiments and Analysis

We give a visualized example to compare our experimental results with real-life events. When a user typed “Microsoft”, our system could detect and show the topics of similar trends, such as “Tweetdeck” illustrated in figure 8, to help people better understand his interested topics. Here, ‘Microsoft’ and ‘Tweetdeck’ have both similar temporal dynamics and similar regional distributions in January and February, 2009. Through our survey, we find out that in early 2009, Microsoft released a test version of windows 7, but users complained that Tweetdeck (an application) cannot run properly on the windows 7, which triggered a hot discussion. We continue to follow these

two topics, and find Microsoft released a concept plan of application 'Next-Generation Newspaper', which is somewhat similar with Tweetdeck, in September. This event also caused hot discussions, leading to another high similarity between these two topics from mid-September to the early October. Due to space constraints of this paper, we don't give the corresponding curves.

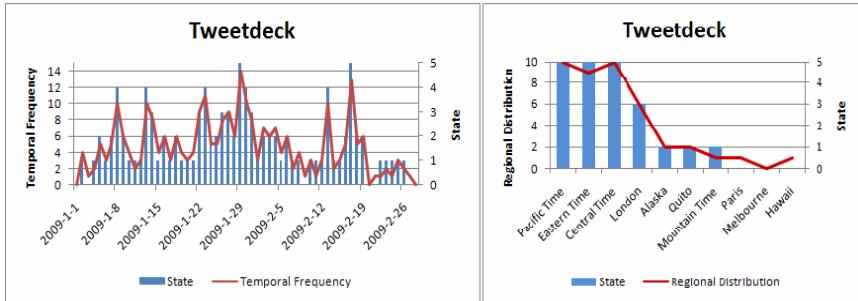


Fig. 8. State series of 'Tweetdeck'

Table 1. Topics related to 'Kobe' in 2009

Kobe	The Possible Cause
NBA	National Basketball Association (NBA), USA.
LeBron	James's name, Kobe's opponent.
Orlando	Refer to Magic, one of the teams in NBA.
Lakers	Kobe's team.
Conan	An TV anchor, Kobe attended his TV show.
Memorial Day	There were many tweets about Kobe losing games around the Memorial Day, May 25, 2009.
Father's Day	Lakers won the championship on the same day, perfect gift for Kobe.
Cavs	LeBron's team, Kobe's opponent.
Kris Allen	'American Idol' Kris Allen to sing national anthem at Lakers game Sunday, June 7, 2009.
Letterman	A word to describe a good player.

Another interesting example we find in twitter topics is about Kobe, a famous basketball player in National Basketball Association (NBA), USA. After applying our spatio-temporal method on the topics which appeared more than 100 times during 2009, 10 most relevant topics to 'Kobe' are discovered, which are shown in table 1.

It can be found from table 1 that almost all of these topics have clear links with Kobe, and others have implicit associations with him, like the "Father's Day" and "Kris Allen". These topics contain his team, his opponents, other teams in NBA and the TV anchor whose program Kobe participated in. For the users who are interested in Kobe, those related topics can not only increase their understanding of Kobe, but also find some interesting anecdotes about him.

Accordingly, in table 2, we list the top ten topics related to 'Kobe' in every month, 2009. In table 2, the number after each month means the frequency of 'Kobe' turned

Table 2. Topics related to ‘Kobe’ in every month, 2009

	January (5)	February (18)	March (7)	April (18)
Kobe	Hulu	NBA	Coraline	Lakers
	USA	World Series	Google Voice	BBQ
	Bing	Jimmy Fallon	LeBron	NBA
	Michael Jackson	Bing	Santa	Follow Friday
	Cowboys	Yankees	#nowplaying	Rihanna
	Spring	Inauguration	Superbowl	Earth Hour
	MySpace	Beyonce	Arkham Asylum	Church
	LeBron	Megan Fox	Vampire Diaries	New Moon
	Iran	White House	#teaparty	CES
	NBA	Follow Friday	Slumdog Millionaire	#MM
	May (117)	June (121)	July (7)	August (0)
	Lakers	NBA	NBA	
	Google Wave	True Blood	Cavs	
	LeBron	LeBron	Megan Fox	
	NBA	Lakers	Lakers	
	California	Lost	US Open	
	MTV	Michael Vick	Kris Allen	
	BBQ	Orlando	Santa	
	Dodgers	Easter	ODST	
	Orlando	CES	Zombieland	
	Texas	Fridays	SXSW	
	September (0)	October (4)	November (8)	December (33)
	Celtics	Miami	Italy	
	Eminem	Celtics	Lakers	
	White House	England	Florida	
	Easter	Chris Brown	LeBron	
	LeBron	Conan	Paris	
	Thanksgiving	BBQ	Cavs	
	Spring	Harry Potter	Michael Jackson	
	Lost	District 9	BBQ	
	Drake	Summer	Halloween	
	Black Friday	Slumdog Millionaire	Swine Flu	

up in the month. Distinguishingly, there is no related record of Kobe in August and September. Compared to the result in table 1, the difference is that the accuracy of the related topics we find in the table 2 is lower. Comparing those results in different months with each other, we can also find that in May and June, the accuracy is higher than that in other months. So, we can deduce some conclusions that: 1. The longer time we compare two topics with our spatio-temporal model, the higher accuracy we can get; 2. We can also get a high accuracy in the burst period of those topics.

The above two examples have given a simple description of our experimental results, through which we can see that our proposed method can effectively detect the correlations among topics in micro-blogging, and through these associations, some interesting web contents can be presented to micro-bloggers.

5 Conclusions and Future Work

In this paper we formulate the task of mining the potential correlations among topics in micro-blogs as a problem of detecting the similar spatio-temporal state series. From

the experimental results, we can see that our similarity-based searching method can effectively discover potential correlations among topics in micro-blogs. This similarity-based method can also be used to do the research of ‘query expansion’, ‘topics recommendation’, and ‘Time series clustering’, etc.

According to the analysis in section 4, we plan to add the content of burst detection prior to the detection of related topics. Future work also includes building the Group-Topic model in Twitter to mine the groups of topics which have intense correlations.

Acknowledgments. This research is partly supported by the projects 863 (No. 2006AA010106), 973 (No. 2007CB311007), NSFC (No. 60703085).

References

1. Lenhart, A., Fox, S.: Twitter and status updating. Pew Internet & American Life Project (February 2009)
2. Platakis, M., Kotsakos, D., Gunopoulos, D.: Searching for Events in the Blogosphere. In: WWW 2009, pp. 1225–1226 (2009)
3. Shieh, J.R., Hsieh, Y.H., Yeh, Y.T., Su, T.C., Lin, C.Y., Wu, J.L.: Building term suggestion relational graphs from collective intelligence. In: WWW 2009, pp. 1091–1092 (2009)
4. Platakis, M., Kotsakos, D., Gunopoulos, D.: Discovering Hot Topics in the Blogosphere. In: EUREKA 2008, pp. 122–132 (2008)
5. Cui, H., Wen, J., Nie, J., Ma, W.: Probabilistic Query Expansion using Query Logs. In: Proceedings of the 11th International Conference on World Wide Web, pp. 325–332. ACM, New York (2002)
6. Li, Z., Wang, B., Li, M., Ma, W.-Y.: A probabilistic model for retrospective news event detection. In: Proceedings of SIGIR 2005, pp. 106–113 (2005)
7. Mitra, M., Singhal, A., Buckley, C.: Improving Automatic Query Expansion. In: Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1998)
8. Qiu, Y., Frei, H.-P.: Concept based Query Expansion. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1993)
9. The Oprah Winfrey Effect on Twitter, April 21 (2009),
<http://www.labnol.org/internet/oprah-winfrey-effect-on-twitter/8274/>
10. Java, A., Song, X., Finin, T., Tseng, B.: 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, WebKDD/SNA-KDD 2007, pp. 56–65. ACM, New York (2007)
11. Ben-Ari, E.: Twitter: What’s All the Chirping About? BioScience 59(7) (July/August 2009), doi:10.1525/bio.2009.59.7.19.
12. Lu, Z., Agarwal, D., Dhillon, I.S.: A spatio-temporal approach to collaborative filtering. In: RecSys 2009, pp. 13–20 (2009)
13. Mei, Q., Liu, C., Su, H., Zhai, C.X.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: WWW 2006, pp. 533–542 (2006)
14. Huang, S., Wu, X., Bolivar, A.: The effect of title term suggestion on e-commerce sites. In: WIDM 2008, pp. 31–38 (2008)

15. Syeda-Mahmood, T.F., Wang, F., Beymer, D., London, M., Reddy, R.: Characterizing Spatio-temporal Patterns for Disease Discrimination in Cardiac Echo Videos. In: MICCAI (1), pp. 261–269 (2007)
16. Ribeiro-Neto, B., Cristo, M., Golher, P.B., Moura, E.S.d.: Impedance Coupling in Content-targeted Advertising. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2005)
17. Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic Query Expansion Using SMART: TREC 3. In: Overview of the Third Text REtrieval Conference, TREC-3 (1994)
18. Xu, J., Croft, W.B.: Improving the Effectiveness of Information Retrieval with Local Context Analysis. ACM Press, City (2000)
19. Kellt, R.: Twitter Study - August 2009. In: Twitter Study Reveals Interesting Results About Usage. Pear Analytics, San Antonio (2009)
20. Gligorov, R.R., Aleksovski, Z., Kate, W.T., Harmelen, F.V.: Using Google Distance to Weight Approximate Ontology Matches. In: Proc. Int'l Conf. World Wide Web 2007, pp. 767–776 (2007)
21. Han, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco (2005)