Employing Twitter Hashtags and Linked Data to Suggest Trending Resources in a Digital Library

Ioannis Papadakis[1], Konstantinos Kyprianos[1(✉)], Apostolos Karalis[2], and Christos Douligeris[2]

[1] Department of Archives, Library Science and Museum Studies, Ionian University, Ioannou Theotoki 72, 49100 Corfu, Greece
papadakis@ionio.gr, k.kyprianos@gmail.com
[2] Department of Informatics, University of Piraeus, Karaoli & Dimitriou 80, 18534 Piraeus, Greece
{akaralis,cdoulig}@unipi.gr

Abstract. It is common truth that social web sites have dominated the web during the past few years. This results in the creation of vast amounts of information that is being produced by the corresponding user activities. Traditional information organization tools originating from the library domain are not applicable to the social web due to its overwhelmingly dynamic nature. Along these lines, hashtags have become an information organization tool of growing popularity among social web sites.

In this paper, it is argued that digital libraries may exploit information deriving from hashtags bringing this way two fundamentally different worlds closer to each other. Thus, a methodology is proposed, where popular hashtags are expanded through semantic web technologies and are ultimately matched against the subject index of a digital library. Successful matches are promoted to the homepage of the digital library to suggest trending resources to the end-users.

Keywords: Social web · Twitter · Hashtags · DBpedia · Semantic web · Linked data

1 Introduction

Digital libraries and libraries in general, are traditionally interested in providing up-to-date information to their users. Thus, libraries most frequently analyse usage statistics [16] in an effort to support their decisions in collection development. Such decisions concern the future of the underlying collections and are based on transactions that occurred in the past.

During the past few years, the social web has emerged as a communication channel capable of facilitating instant information sharing and collaboration. One of the most widespread features of the social web are the hashtags. Hashtags are employed by social web tools to classify messages, propagate ideas and promote specific topics and people [8]. They have been introduced by Twitter, but nowadays are commonly met in many social web tools. Hashtags have evolved into a powerful

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 407–418, 2017.
DOI: 10.1007/978-3-319-67008-9_32

classification tool for the social web. They constitute a peculiar kind of dynamic vocabulary that is controlled by the same people that employ it (i.e. the end-users). Hashtags are searchable through Twitter, Google and specialized sites [18].

In this paper, a methodology is proposed that evolves around hashtags and suggests resources within a digital library about trending topics. In contrast to traditional library practices, the proposed approach aims in observing current trends in the society and instantly suggesting relevant resources to its users. The proposed methodology is implemented and deployed as a service within the context of an academic digital library. Finally, the service is evaluated and interesting remarks are shown.

The remainder of this paper is structured as follows: In the following section, a short description is presented about the relation between social networks, Twitter, hashtags and digital libraries. Then, related research regarding the employment of hashtags across the social web is mentioned. Next, an effort is made to highlight the convergence and divergence points between controlled vocabularies and hashtags. In Sect. 5, the proposed methodology is realized as an online service and its deployment in a digital library is presented. In the following section, the service is accordingly assessed and the corresponding results are shown. Finally, Sect. 7 concludes the paper and points directions for future work.

2 Social Networks, Twitter, Hashtags, Digital Libraries

Since their initial appearance, Social Network Sites (SNS), such as Facebook, Twitter, and Myspace, have attracted millions of users globally. In fact, many users have rendered such sites an integral part of their lives [4]. The widespread of collaborative technologies has led to the formulation of instant online communities, thus facilitating the communication among people rapidly and conveniently [10].

Twitter is a microblogging service that allows users to follow other users or to be followed. Unlike most other SNS, the relation between following and being followed is not symmetric. A user may follow other users but it is not necessary to be the other way around [13]. Twitter allows users to broadcast brief text updates about things that are happening to their lives. Users refer to Twitter when they want to find information about breaking news, real-time events, people and topical information [14, 19]. Such features establish Twitter as a tool that may provide timely information quicker than any other mass media (e.g. television, radio, etc.). According to January 2017 report by Statista[1], Twitter is among the 10 most famous SNS, having more than 317 million active users. Moreover, based on Alexa website ranking[2], Twitter possesses the 16th place in the global rank regarding the most famous sites across the world.

One popular feature of Twitter is the employment of hashtags. A hashtag is a convention among Twitter users to create and follow a thread of discussion by prefixing

[1] Most famous social network sites worldwide as of January 2017, r. (2017). *Global social media ranking 2017 | Statistic*. *Statista*. Retrieved 28 March 2017, from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[2] *Keyword Research, Competitor Analysis, & Website Ranking | Alexa*. (2017). *Alexa Internet*. Retrieved 28 March 2017, from http://www.alexa.com.

a word with a '#' character [13]. Tweets containing a hashtag are visible not only to the followers of a user that employed the specific hashtag but to anyone on the social network. Twitter provides an API through which it is possible to identify the most popular hashtags so that users can witness trending topics. A trending topic does not last forever nor does it disappear so as to never come back. Usually, it lasts for a couple of days and can have many active periods [13].

Thus, hashtags, to some extent, reflect the trending topics that users are talking about. Through the employment of hashtags, tweets can be organized, indexed, shared and discovered by anyone [12]. Consequently, hashtags can be seen as a powerful vocabulary that is created, employed and controlled by the users themselves.

In this paper, it is argued that digital libraries may exploit information deriving from hashtags in favor of their users by pinpointing trending resources that exist in their collections. Thus, digital libraries get valuable feedback from external entities in a timely fashion.

3 Using Hashtags Across the Social Web

As already mentioned, since their emergence on Twitter[3], hashtags have been used extensively in social networks and micro-blogging services. Their wide acceptance urged many researchers to study them thoroughly.

Efron [9] proposed a language modelling approach to hashtag retrieval based on the assumption that when a user is interested in a specific topic he/she might like to find hashtags that are often applied to posts about such topic. In a similar fashion, Bansal, Jain and Varma [3] proposed a method of semantic enrichment of microblogs for a particular type of entity search that ends up in retrieving a ranked list of the top-k hashtags relevant to a user's query. Such a methodology may help users to track posts of general interest. The aforementioned approaches aim in providing end users with personalized information from tweets based on hashtags.

In a different approach, Sedhai and Sun [17] introduced an entity-hashtag graph for tweets with hyperlinks. More specifically, they grouped together the hashtags of tweets containing links to various web pages and recommended them to future users that posted tweets containing the same web pages. Thus, information sharing and organization within the Twitter ecosystem can be facilitated.

Another indicative case of research work about hashtags that evolves around the Twitter ecosystem is TweetPos, which was proposed by Wijants et al. [21]. TweetPos is a versatile web-based tool that facilitates the analytical study of geographic tendencies in crowd-sourced Twitter data feeds. Hashtags play a crucial role in this tool, since they constitute the service's essential ingress parameters. When a user addresses a topic query to TweetPos, the system creates a compilation of tweets about this topic. The user may geographically and temporally filter such tweets.

[3] *#OriginStory - Carnegie Mellon University.* (2017) *#OriginStory.* Retrieved 29 March 2017, from http://www.cmu.edu/homepage/computing/2014/summer/originstory.shtml.

In a slightly different line of research, the following two approaches exploit hashtags to retrieve real-time and popular events that people are discussing in Twitter. More specifically, Wang et al. [20] proposed an adaptive crawling model that detects emerging popular hashtags and monitors them to retrieve high volume of relevant data for events of interest. The model analyzes the traffic patterns of the collected hashtags to update subsequent collection queries. Cui et al. [7] aim in discovering breaking events with the employment of popular hashtags in Twitter.

There are also some approaches that facilitate the search and retrieval of topic-related tweets with the employment of hashtags. More specifically, the methodology proposed by Llewellyn et al. [15] focuses on the formulation of a corpus of tweets about a specific topic based on popular hashtags, hand-selected hashtags and topic modelling. In a similar approach, Cotelo, Cruz and Troyano [6] proposed a general, dynamic and graph-based model to capture related but unknown topics in tweets based on hashtags and users. Bansal, Bansal and Varma [2] presented a machine learning methodology to segment the hashtags and link the entities in hashtags to Wikipedia, an approach that helps in finding latent semantic information about hashtags.

To conclude, it seems that a great deal of the relevant literature about hashtags emphasizes on finding ways to aid users in discovering additional hashtags and tweets regarding their initial information needs. Consequently, the rich semantic information that lies into hashtags is mostly exploited within the strict boundaries of the Twitter ecosystem.

4 Controlled Vocabularies vs. Hashtags

Controlled vocabularies can be seen as collections of terms defined by experts that are employed to index and, ultimately, to retrieve information through browsing or searching [11]. Controlled vocabularies typically include preferred, non-preferred and related terms. In many cases, these terms have hierarchical relationships among them, meaning that navigation is possible from a generic term to a more specific one and vice-versa. The purpose of controlled vocabularies is the organization of information and the provision of terminology to catalogue and retrieve information [11].

On the other hand, hashtags can be seen as vocabularies that are defined by common users and not by experts [5]. Moreover, instead of referring to formal collections, hashtags are employed to index messages or tweets on the microblog sphere. Such terms are incorporated into a tweet by the author of the specific tweet, meaning that there is no limitation or control over the term that will be created as a hashtag. After their publication, the decisive factor that transforms a hashtag to something like a controlled vocabulary term is popularity. Popular hashtags tend to be employed in many tweets thus becoming even more popular. Therefore, it seems that hashtags go under a constant control and evaluation by the users themselves. A user that decides to employ an existing hashtag, promotes this hashtag and ultimately contributes to its establishment as an authority. To sum up, it is evident that formal controlled vocabularies created by experts and popular hashtags created by users have signs of convergence.

5 Proposed Approach

In this section, we propose a digital library service, which harvests trending hashtags from Twitter to identify relevant resources within a digital library. Such resources are promoted to the homepage and then suggested to end-users through an interactive query suggestion service. The service is based on a technique that was introduced at [16].

Initially, the most popular tweets are harvested and their corresponding hashtags are stored for further process. Since hashtags rarely look like normal terms that usually exist within library indices, such hashtags undergo a spell-checking control. The controlled hashtags are tunneled towards DBpedia [1], a linked data provider containing structured content deriving from Wikipedia, in an effort to enrich the term collection with even more relevant terms. The structured information can be queried online through the employment of appropriate semantic web technologies (e.g. SPARQL). The enriched set of terms originating from popular tweets is matched against the subject index of a digital library. The successful matches are again stored and the more recent ones are ranked at the top. The top-n terms populate an HTML division element (i.e.) at the homepage of a digital library. The end-users are able to interact with such terms and accordingly retrieve resources from the digital library about trending topics on Twitter.

The next section provides a detailed analysis of how the proposed service offers the aforementioned functionality.

5.1 Implementation Details

To suggest resources about trending topics in the digital library, the underlying engine goes incrementally through the phases below (see Fig. 1)[4]:

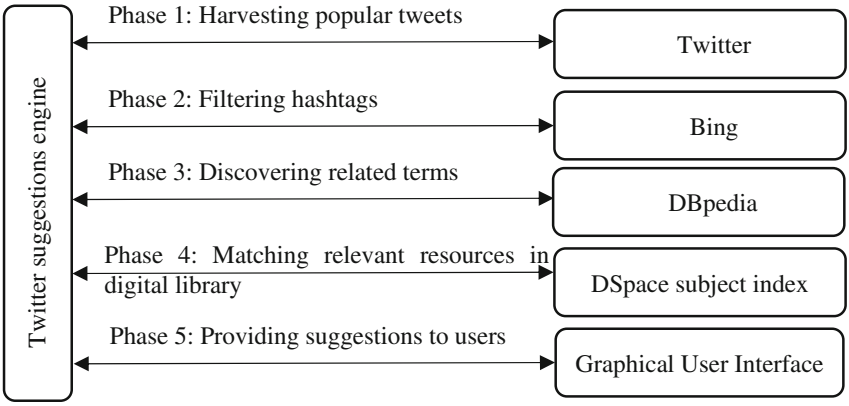

Fig. 1. Information workflow

[4] The source code of the service is available at bitbucket: https://bitbucket.org/akaralis/twitter-inlibraries.

Phase 1: The *n* most popular tweets are harvested in Twitter. For this purpose, Twitter's[5] Search API is used.

Phase 2: The hashtags of each tweet are filtered using a spelling suggestion service, namely the Bing Spell Check API[6]. For example, the phrase "Cloud Computing" is produced when the hashtag #CloudComputing is filtered using the aforementioned spelling suggestion service.

Phase 3: For the resulting set of spelling suggestions and the unfiltered hashtags, related terms and categories are requested via the SPARQL endpoint of DBpedia[7].

Phase 4: The entire set of the spelling suggestions, the unfiltered hashtags and the related terms and categories are matched against the subject index of a digital library. Obviously, the successful matches of this phase constitute the set of trending topics that are ultimately suggested to end-users.

Phase 5: Each time a client (i.e. a browser on which the homepage of the digital library is loaded) requests suggestions, the proposed service retrieves from the underlying database and returns a predefined number of the most recent ones that have the highest number of "clicks" (by the users) in a specified timeframe (e.g. the last 5 days). This way, the service promotes recent suggestions that are popular among the population of the digital library.

5.2 Deployment of the Proposed Service

The proposed service is implemented and deployed in Dione, the academic digital library of the University of Piraeus in Greece[8]. Dione mainly contains theses and dissertations from students of the four Schools of the University (i.e. School of Economics, Business and International Studies, School of Finance and Statistics, School of Maritime and Industrial Studies and School of Information and Communication Technologies).

The initial goal for the service was to suggest queries that would apply to all the scientific disciplines that are relevant to the Schools of the University. However, it was soon realized that popular hashtags contained too much noise and ambiguity that led to meaningless suggestions. Therefore, it was decided to filter the initial tweets and select just the ones that contain the word 'programming' to focus on tweets about technology.

The alpha version of the service was attached to Dione in January 2016 and the beta version was attached to the digital library on January 15th, 2017. The beta version introduced a new algorithm for populating the HTML division element of Dione's homepage. During the alpha version, the five most recent terms of the database appeared in the

[5] *The Search API*. (2017). *Twitter Developer Documentation*. Retrieved 20 April 2017, from https://dev.twitter.com/rest/public/search. More specifically, the parameter result_type='mixed' is employed in an effort to get a mix of recent and popular tweets.

[6] *Bing Search API | Microsoft Azure Marketplace*. (2017). *Datamarket.azure.com*. Retrieved 20 April 2017, from http://datamarket.azure.com/dataset/bing/search.

[7] The corresponding SPARQL query would be: SELECT ?related WHERE {<http://dbpedia.org/resource/Data_science><http://purl.org/dc/terms/subject>?related}. DSpace REST API, avail. at: https://wiki.duraspace.org/display/DSDOC5x/REST+API [accessed: 12/03/2017].

[8] *Dione* (2017). *Dione's Homepage*. Retrieved 29 March 2017, from http://dione.lib.unipi.gr.

division element. When a new term emerged in the database, it took the place of the oldest one in the division element. In the beta version, the division element contains seven terms. Each time a user selects a term, a counter is increased. When a new term emerges in the database, it substitutes the last term of the division element, which contains terms ranked by both the number of times they have been selected and their age. No term can stay in the division element more than five days.

As shown in Fig. 2, the proposed service is visualized as a division HTML element (i.e.) containing suggested queries about trending topics at the top right of the digital library's homepage. Upon selection, a query is addressed to the underlying search engine and the matched resources are returned to the user.

Fig. 2. Dione's homepage

6 Evaluation

The proposed service has been evaluated in order to assess the impact to the user community of the digital library as well as the performance of the various modules that constitute the service. The evaluation is based on a log file analysis of Dione's usage from January 15th, 2017 until March 15th, 2017.

6.1 Impact to the User Community

The log files of Dione provide the opportunity to compare the number of times the proposed service has been employed to the number of times the rest of the six browsing interactions of Dione have been used by the community. As shown in Fig. 3, the most

popular navigational interaction is against the subject index (61.07%), followed by the author index (16.2%). The advisor index ranks third (15.44%) and Twitter suggestions appear in the fourth place (4.75%). The remaining interactions are not very popular since they cover less than 2% of the total number.

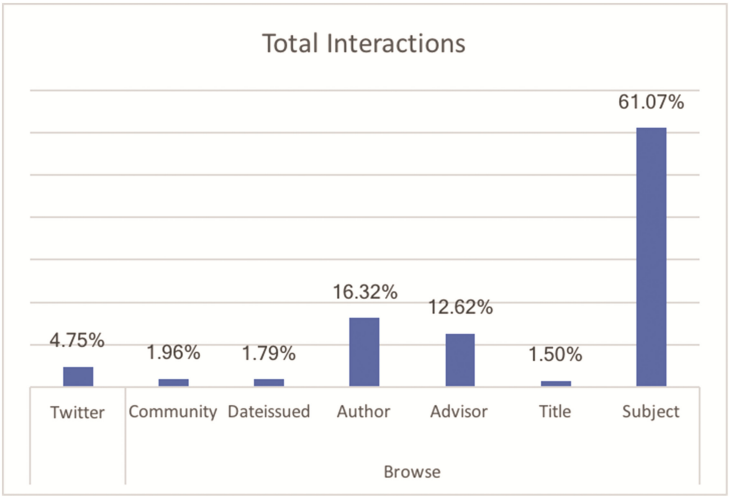

Fig. 3. Total Interactions

It is apparent that the number of Twitter suggestions is significantly lower than the top-three interactions (i.e. 'Subject', 'Author' and 'Advisor') provided by Dione. However, the proposed service ranks higher than the rest of the interactions that lie below. This is particularly important if one takes into consideration the fact that the scope of Twitter suggestions is limited to technology (through the employment of the word 'programming'), whereas the other interactions refer to the total of the disciplines covered by Dione.

Moreover, 22.9% of all the distinct Twitter suggestions that appeared in the element were actually selected by the user community (63 out of 275 distinct suggestions). Thus, to a great extent, the service did not manage to filter out terms that went later unnoticed by the users of Dione. This could be attributed to the fact that too much Twitter noise managed to penetrate into the proposed service.

Since the business logic of the proposed service dictates that each suggestion may appear in the element of the homepage once for each five days' timeframe, it becomes apparent that a single suggestion may appear in the element many times in different timeframes. Thus, it would be interesting to see whether the number of times each hashtag appears in the element influences the number of times the hashtag is selected by the users. According to Fig. 4, hashtags that make it often to the element have a higher chance of being selected by users. Such a conclusion certainly comes as no surprise, since frequently appearing hashtags have more chances to get selected. Thus, it seems that the users of Dione follow the trending topics as they have been recorded in Twitter.

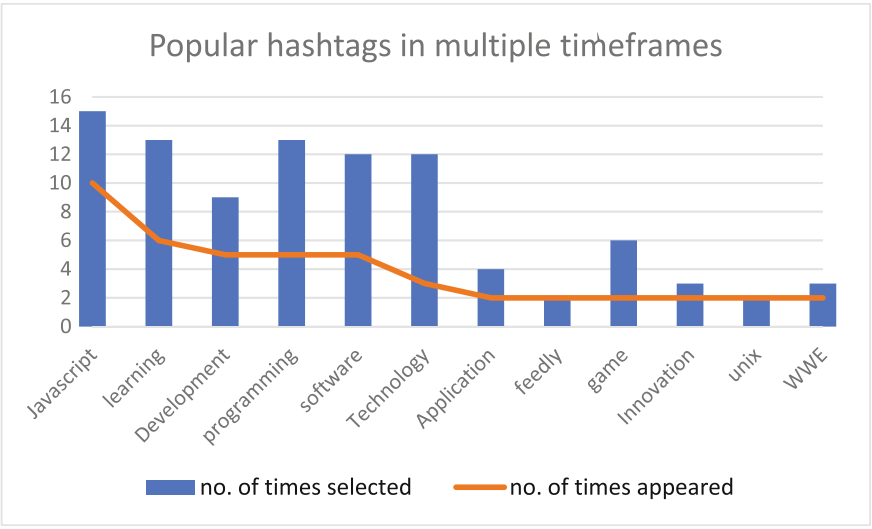

Fig. 4. Popular hashtags that have been selected by users in multiple timeframes

6.2 Under the Hood: Performance of Individual Modules

In the previous section, the impact of the proposed service to the user community of Dione was assessed. In this section, further analysis is performed to assess the various interactions between the core modules of the service and the remote online services that have been employed (namely: Twitter, Bing and DBpedia).

The vast majority (i.e. 71%) of the keywords that made it to the element of the homepage come from DBpedia, whereas 17% of them originate from Twitter alone and just 2% have taken advantage of Bing's spelling suggestion service as seen in

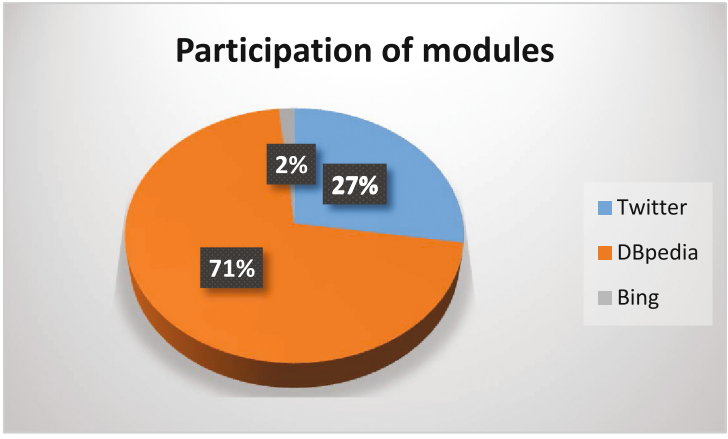

Fig. 5. The impact of the three remote modules that are employed by the proposed service

Fig. 5. It is apparent that DBpedia is very influential to the proposed service since it manages to provide useful keywords to the user community. On the other hand, Bing's spelling suggestion service has a minimal effect on the service.

From another point of view, it would be interesting to assess the impact of the three modules to the users of Dione. Thus, Fig. 6 calculates the number of times each suggestion has been selected from Dione's population, grouped by the module each suggestion originates from. This time, it is apparent that suggestions from Twitter are more popular than suggestions from DBpedia, despite the fact that most of them originate from DBpedia.

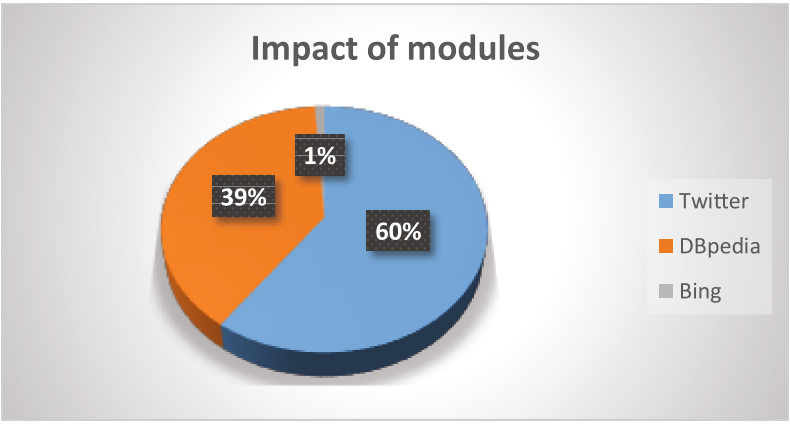

Fig. 6. The impact of the three modules to the users of Dione

7 Conclusions and Future Work

In this paper, a methodology was proposed that takes advantage of information that exists outside a digital library in favor of a query suggestion service within a digital library. The service suggests resources about trending topics that have been harvested from Twitter and expanded through DBpedia. The proposed service was integrated in the institutional repository of an academic library and was thoroughly evaluated.

The assessment process was based on quantitative methods. A log file analysis indicates that the proposed service attracts a considerable number of digital library users. Moreover, further analysis on the functionality of the service shows that the employment of DBpedia improves both the quantity and the quality of the provided term suggestions. Future work focuses on minimizing the inherent "noise" of the hashtags and on finding ways of applying the proposed approach to the entire scope of the digital library.

This work is a first step towards the integration of traditional digital library services with information originating from popular, crowd-sourcing sites that exist on the Web. The evaluation of the proposed service exhibits promising results in terms of the appreciation of the service by the users of the digital library. At the same time, it is evident that the lack of control to the information that is being accumulated in such sites

contradicts the strict organization principles that traditionally govern digital libraries. Therefore, further detailed analysis is required prior to the entrance of such information to any digital library. The semantic web technologies could play a crucial role in minimizing the inherent noise of crowd-sourced information and become the catalyst for the creation of new, value-added services in the digital library domain.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). doi: 10.1007/978-3-540-76298-0_52
2. Bansal, P., Bansal, R., Varma, V.: Towards deep semantic analysis of hashtags. In: Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.) ECIR 2015. LNCS, vol. 9022, pp. 453–464. Springer, Cham (2015). doi:10.1007/978-3-319-16354-3_50
3. Bansal, P., Jain, S., Varma, V.: Towards semantic retrieval of hashtags in microblogs. In: Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May, pp. 7–8 (2015)
4. Boyd, D.M., Ellison, N.B.: Social network sites: definition, history, and scholarship. J. Comput.-Mediated Commun. **13**(1), 210–230 (2007)
5. Bradley, P.: Social Media for Creative Libraries, 2nd edn. Facet Publishing, London (2015)
6. Cotelo, J.M., Cruz, F.L., Troyano, J.A.: Dynamic topic-related tweet retrieval. J. Assoc. Inf. Sci. Technol. **65**(3), 513–523 (2014)
7. Cui, A., Zhang, M., Liu, Y., Ma, S., Zhang, K.: Discover breaking events with popular hashtags in Twitter. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), Maui, HI, 29 October–02 November, pp. 1794–1798 (2012)
8. Cunha, E., Magno, G., Comarcel, G., Almeida, V. Goncalves, M.A., Benevenuto, F.: Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach. In: Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 58–65 (2011)
9. Efron, M.: Hashtag retrieval in a microblogging environment. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland, 19–23 July, pp. 787–788 (2010)
10. Fu, F., Liu, L., Wang, L.: Empirical analysis of online social networks in the age of web 2.0. Phys. A: Stat. Mech. Appl. **387**(2–3), 675–684 (2008)
11. Harpring, P.: Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works. Getty Research Institute Publications, Los Angeles (2010)
12. Ivankova, M.: Understanding Microblogging Hashtags for Learning Enhancement. Form@re **11**(74), 17–23 (2011)
13. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, 26–30 April, pp. 591–600 (2010)
14. Lau, C.H., Li, Y., Tjondronegoro, D.: Microblog retrieval using topical features and query expansion. In: Text REtrieval Conference (2011)
15. Llewellyn, C., Grover, C., Alex, B., Oberlander, J., Tobin, R.: Extracting a topic specific dataset from a twitter archive. In: Kapidakis, S., Mazurek, C., Werla, M. (eds.) TPDL 2015. LNCS, vol. 9316, pp. 364–367. Springer, Cham (2015). doi:10.1007/978-3-319-24592-8_36

16. Papadakis, I., Kyprianos, K., Karalis, A.: Highlighting timely information in libraries through social and semantic web technologies. In: Garoufallou, E., Subirats Coll, I., Stellato, A., Greenberg, J. (eds.) MTSR 2016. CCIS, vol. 672, pp. 297–308. Springer, Cham (2016). doi: 10.1007/978-3-319-49157-8_26
17. Sedhai, S., Sun, A.: Hashtag Recommendation for hyperlinked tweets. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, Queensland, Australia, 06–11 July, pp. 831–834 (2014)
18. Small, T.A.: What the hashtag?: a content analysis of canadian politics on Twitter. Inf. Commun. Soc. **14**(6), 872–895 (2011)
19. Teevan, J., Ramage, D., Morris, M.R.: #TwitterSearch: a comparison of microblog search and web search. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM 2011), Hong Kong, 9–12 February, pp. 35–44 (2011)
20. Wang, X., Tokarchuk, L., Cuadrado, F., Poslad, S.: Exploiting hashtags for adaptive microblog crawling. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, Canada, 25–28 August, pp. 311–315 (2013)
21. Wijnants, M., Blazejczak, A., Quax, P., Lamotte, W.: Geo-spatial trend detection through Twitter data feed mining. In: Monfort, V., Krempels, K.-H. (eds.) WEBIST 2014. LNBIP, vol. 226, pp. 212–227. Springer, Cham (2015). doi:10.1007/978-3-319-27030-2_14