Socio-semantic query expansion using Twitter hashtags

Ioannis Anagnostopoulos
Dpt. of Computer Sciences and Biomedical Informatics
University of Central Greece
Lamia, Greece
janag@ucg.gr

Vasileios Kolias, Phivos Mylonas
School of Electrical and Computer Engineering
National Technical University of Athens
Zographou Campus - Athens, Greece
{vkolias@medialab, fmylonas@image}.ntua.gr

***Abstract*— In this paper we introduce an algorithmic approach, capable of creating a semantic network with concatenated terms and phrases (hashtags) from collectively postings on the Twittersphere. This network could be exploited for query expansion provision in respect to users' information needs, without considering any other prior knowledge or access in search logs or browsing history records. For evaluation purposes, we compare our query expansion approach algorithm with query suggestions provided by well-known search engines and mainstream media services (e.g. Google, Yahoo!, Bing, NBC and Reuters), as well as by enrolling a team of human editors, who provided subjective comparisons in respect to the Google Hot Searches service. The results are quite promising, showing that our proposal semantically expands the user's initial query with related terms adapted to social trends and knowledge.**

Keywords - query expansion, social media, Twitter hashtags, semantic network

I. INTRODUCTION

Nowadays, the amount of information disseminated on the Web is without doubt enormous. Thus, the Web itself, its users, their user queries and Web search engines form a gigantic system that exchanges and circulates data. Thus, modern information services provide a lot of mechanisms for suggestions in respect to users' information needs expressed by mostly syntactic queries. Research on query suggestion is highly related with query expansion [1], query substitution [2], query recommendation [3] or query refinement [4]. All are considered as similar procedures aiming to adjust an initial user query into a revised one, which then returns more accurate results. However, query suggestion is closely related to query analysis, as the query, the user and the medium that transfers it have to be examined.

Entities (i.e. factors and actors) involved in query analysis consist typically of users, their queries, information systems (e.g. search engines) and the Web it self. Since all of them are interconnected, one may further extend this classification, resulting in effective improvements of existing approaches, as well as novel methodologies. The classification derived from above entities is depicted in Figure 1. For example, users tend to adapt their queries to the capabilities of search engines (SE), while being affected from the information they get. On the other hand, websites also tend to adapt to search engines, to acquire the highest rank possible in the returned search results. Even the search engines themselves tend to adapt to the human ways of thinking, reasoning and communicating in order to increase their effectiveness. All the aforementioned factors provide different dynamics to query analysis, thus making the constant re-evaluation of methodologies and their results a necessity. In this work, we deviate from the traditional query suggestion proposal (as depicted in Figure 1) in a sense that users have their queries expanded directly from social media platforms (in this work we use Twitter[1]), and without having their queries or browsing history processed by search engines (Social Media-to-User Suggestions).

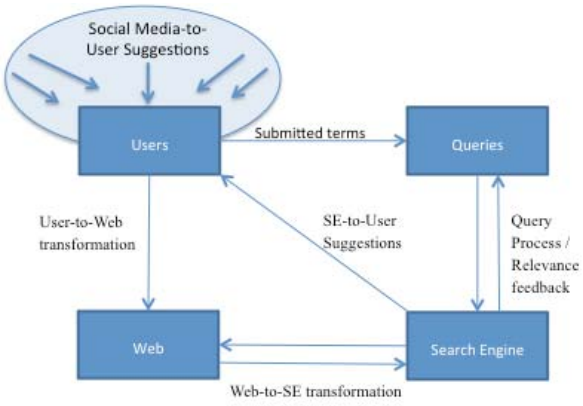

Figure 1. Constituent actors in query analysis

The remainder of this paper is organized as follows. Section II provides the methodology we use, as well as the basic steps of our proposed algorithm. In this section we also describe a case study in order to clearly show how our query expansion mechanism works. Section III has a two-fold role. It firstly presents an experimentation conducted in order to evaluate the case study results, while it further discuss an evaluation procedure, which enrolled seven human editors (raters) who provided subjective comparisons between the results derived from our algorithm and Google. Finally, this work ends up with the conclusions and our thoughts for future work in Section IV.

II. PROPOSED METHODOLOGY

In this section we briefly present the third-party information source we use, as well as the basic algorithmic steps that aid us to introduce social and semantic knowledge

[1] http://www.twitter.com

978-0-7695-4887-6/12 $26.00 © 2012 IEEE
DOI 10.1109/SMAP.2012.15

for query expansion with up-to-date terms, with respect to the actual users' information needs.

A. Third-party information sources used

The query analysis and expansion algorithm we utilize is exclusively based on words or concatenated terms and phrases, known as hashtags. These words are prefixed with the symbol "#", which is a form of metadata tag[2]. This type of information is particularly used on Twitter, when users either want to emphasize a term/phrase or intend to aid the social network organize the huge amount of information disseminated globally. Thus, if a hashtag becomes extremely popular, it will appear in the "Trending Topics" area of a user's homepage. There are no strict syntactic rules (apart from the concatenated form in subsequent terms), so users may add/post any word that best represents a concept according to their opinion or knowledge. Thus, on the one hand, users are free to semantically express themselves at will and add knowledge to the social network, but on the other hand, a single hashtag may be used for various unrelated different topics created by those who make use of them.

As a result, the main scope of this paper is to correlate hashtags that correspond to semantically similar topics and are presented in related tweets, in order to use them for query expansion. For this purpose, we use hashonomy[3] as our (third-party) information source, which, in brief, organizes in real-time all links, sources, hashtags and users with respect to the Twitter social network, offering also suggestions to registered users.

B. Algorithm description

In the following we present how we use an algorithmic procedure in order to create a semantic social network containing the trendiest terms with respect to a user's query. This network will be dynamic in nature, including up-to-date information and high impact in terms of social-driven knowledge.

Initially, upon a user's query submission, we extract its query terms (seeds) and then (by utilizing corresponding API) we get the top-k results within a specific period p as provided by the trending timeline of hashonomy. These results are actually clustered tweets that contain the seed(s) in hashtag format and in order of freshness. Then, for all selected hashtags we calculate their Twitter Semantic Weight *TSW(seed→ht)* from the seed, in terms of three separated weighting factors, namely the amount of *Clustered Tweets* related to the seed *CT(ht)*, the *Hashtag Relation* weighting factor *HR(ht, seed)* and the *Period Appearance* weighting factor *PA(ht,p)*, as depicted in Equation 1. *HR(ht)* and *PA(ht)* correspond to how strongly the examined hashtag is related to the seed term, as well as how frequently the hashtag appears within the time interval under investigation.

$$TSW(seed \rightarrow ht) = CT(ht) * HR(seed, ht) * PA(ht,p) \quad (1)$$

where:

HR(seed, ht)=1/(n+1), hashtag co-appears with other n hashtags and not with the seed

HR(seed, ht)=2/(n+1), hashtag co-appears with other n hashtags and with the seed

PA(ht, p)=a/p, *a:* amount of time units the hashtag appeared *(a=1, ..., p)*

After having the scores calculated for all examined hashtags, we select the l-highest seed to hashtag weighting values *TSW(seed→ht)* and then we follow an iterative procedure, setting the examined hashtag as a new seed. Finally, we set the amount of subsequent applications of Equation 1 as the irritation depth d of the algorithm for further semantic investigation between the seed and its related hashtags (provided by hashonomy). In other words, the semantic network derived from the initial seed (query term) to examined hashtags 1,2, …, l has a depth value equal to *0,* while the semantic network derived from hashtag i (i.e. the *new* seed), where *i=1,2, ..., l* to every other related hashtag, has a depth value equal to 1, and so on. At this point it should be pointed out that all above-defined parameters may be fine-tuned, since they strongly affect the size of the social semantic network created and as a result the overall response time of the query expansion procedure.

In the next section, we present a case study of the proposed algorithm. We would like to note here that tokenization, topic/word segmentation, as well as further lexical analysis procedures that deal with breaking a stream of text up into words/phrases are not considered to be part of this work. In any case we rely on the fact that suggested (or expanded) query term(s) is/are provided as appeared in hashtag format and it is understood as initially defined by Twitter users. Through our proposed methodology a semantic network of related terms is created directly from users tweets. As it will be presented in the following a social semantic network is dynamically created capable of suggesting related terms to users during their web search.

C. Case Study

As case study we use the repugnant crime that took place on July 20, 2012 in the City of Aurora, Colorado, U.S.A., where a gunman wearing a gas mask firstly set off an unknown gas and then fired into a crowded movie theater during a midnight screening of the Batman film "The Dark Knight Rises", thus killing 12 people and injuring at least 50 others. This shocking event was among the breaking news globally for several days and emotionally touched people worldwide. Apart from mainstream media, social media platforms covered all aspects of the incident disseminating an enormous amount of information, which was created from millions of users. Especially in Twitter, information contained not only shared information, but also personal opinions/thoughts, and constantly new links related to the crime, directly related to user-generated hastags as semantic annotations.

[2] http://en.wikipedia.org/wiki/Hashtag
[3] http://www.hashonomy.com/

So, in this case study we consider the word "*Aurora*" as our initial seed term. It is worth noticing that even after two weeks after the event, most search engines did not suggest relevant terms after the seed term. Our main idea is to provide query expansion to the user's submitted term(s) with publicly posted hashtags directly from Twitter social media and without having any other access or use of search engines' query logs.

Hashtags related to seed (#aurora)			*Metrics*	
#hashtag1	*#hashtag2*	*#hashtag3*	*Number of Tweets*	*Time Interval*
#darknight	#____		56	0
#thedarknightrises	#tdkr	#horor	1	0
#____	#classy		134	0
#theatershooting	#____		295	1
#batman	#____	#violence	1	1
#aurorashootings	#healthcare		291	1
#____	Guns		118	2
#____	#colorado		953	2
#obama	#colorado		132	3
#____	#theatershooting		2K	3
Parameters set				
k = 10	*p = 4*	*d = 0*	*TSW(1)/TSW(l) ≤ 10*	
Hashtags related to seed (#theatershooting)			*Metrics*	
#hashtag1	*#hashtag2*	*#hashtag3*	*Number of Tweets*	*Time Interval*
#batman	#____		122	0
#____	#aurora		116	0
#colorado	#____		52	1
#____	#cofires		12	1
#batman	#joker	#____	51	2
#____	#9news	#journalist	63	2
#jamesholms	#____		14	2
#____	#colorado		57	3
#9news	#jamesholmes		17	3
#darknight	#____		38	3
Parameters set				
k = 10	*p = 4*	*d = 1*	*TSW(1)/TSW(l) ≤ 10*	

Table 1. Example of retrieved hashtags from hashonomy, metrics and related parameters in our case study (#_____ corresponds to the seed in hashtag appearance)

In this sense, Table 1 provides some examples with respect to the proposed methodology. Moreover, in case we seek directly related hashtags for the term "*aurora*" (*d*=0), we utilize the hashonomy API to get the top-10 (*k*=10) clustered results in terms of freshness, thus resulting into a four-day period (*p*=4). The last two columns on the right of Table 1 depict the number of tweets corresponding to a specific cluster of terms, along with the time this information was tweeted according to the hashonomy. For instance, the sequence {#darknight, #_____, 56, 0} defines that hashtags "darknight" and "aurora" appeared together in 56 tweets within the last day, while the sequence {#_____, #theatershooting, 2K, 3} expresses that hashtags "aurora" and "theatershooting" appeared together more than 2000 tweets four days ago.

Having calculated the scoring distances for all top-10 clustered hashtags, we select related hashtags, provided that their *TSW* distance is at least equal to one order of magnitude (*TSW(1) / TSW(l) ⩽ 10)*. As depicted in Figure 2, this rule provides us with an initial semantic network of three related hashtags along with their TSWs (here: #theatershooting/4.62, #colorado/2.27, and #aurorashootings/0.56).

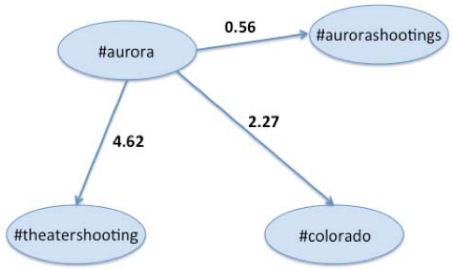

Figure 2. Social Semantic Network (seed: #aurora, k=10, p=4, d=0, TSW_{dist} less than one order of magnitude)

The same applies when we repeat our algorithm setting as seed to the previous related hashtags (d=1 → #theatershooting, #colorado, #aurorashootings). The second part of Table 1 holds the top-10 (k=10) clustered results in terms of freshness for the seed #theatershooting, while p is once more equal to 4. The hashtag sequence {#batman, #_____, 122, 0} defines that hashtags "batman" and "theatershooting" appeared together in 122 tweets within the last day, while the sequence {#darknight, #_____, 38, 3} expresses that hashtags "darknight" and "theatershooting" appeared together in 38 tweets four days ago.

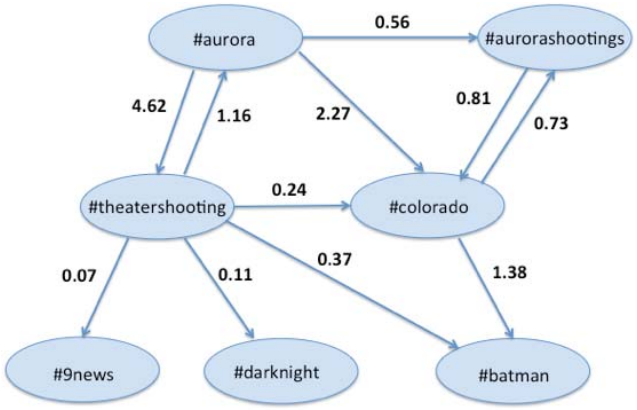

Figure 3. Social Semantic Network (seed: #aurora, k=10, p=4, d=1, TSW_{dist} less than one order of magnitude)

Similarly to above, having set *(TSW(1)/TSW(l)⩽10)* for all top-10 clustered results per seed, we ended up with an extended semantic network derived from the most trendy and highly appeared Twitter hashtags, as illustrated in Figure 3.

III. EVALUATION - DISCUSSION

In order to evaluate our algorithm we provide two separate approaches divided across two following sub-sections. Initially we compare the results derived from our case study to the query suggestions of well-known search engines, like Google, Yahoo!, Bing, as well as mainstream Web media services, such as Reuter News and NBC. In the second sub-section, we describe a generic evaluation, which involves subjective user ratings for results obtained from our query expansion algorithm and Google.

A. Case study evaluation

In this sub-section we provide evaluation of the query expansion and suggestions provided by our algorithm. In order to achieve this, we compare query expansions provided by search engines and mainstream media services, firstly for the term "*aurora*" as well as with other three terms, namely "*theater shooting*", "*colorado*", and "*aurora shootings*". All these terms were derived from our algorithm as seed terms in a hashtag format and were highly relevant with respect to the incident we used within our case study. We also point out that suggestions provided by our algorithm are based exclusively on user-generated hashtags in the twittersphere and not from query logs or other type of information possibly hidden within related information structures. Moreover, Google's predicting algorithm used for query expansion displays search queries based on other users' search activities and the contents of Web pages indexed by Google[4]. In addition Google users might also see search queries from their previous related searches. We suppose that the rest information services we use for comparison purposes do work under the same concepts. Still, we should note at this point, that if the search service uses a *search results* based approach, query expansion depends on a specific number of the top-*N* results for that query. Yet, if the service uses *logs*, query expansion may be provided upon other relevant user query terms, or even other user personalized behavior-based search pattern.

Evaluation results for the semantic social network depicted in Figure 3 are provided within following Table 2. Initial query terms (called seeds) are on the left side of the table, followed by suggested term(s) in a tree form. This resembles to the drop-down list provided by several search engines, which contains suggested terms based on the already submitted user term(s). A column depicting the aggregated TSW follows, which reflects the cumulative TSW of all involved (i.e. selected by the algorithm) terms, divided by their amount/number. Finally, the last four columns of Table 2 indicate whether the specific expanded query has been suggested (even in different order of terms with respect to the seed) by Google, Yahoo!, Bing, NBC and Reuters[5]. The date we performed this evaluation was August 3, 2012, just two weeks after the actual "Aurora incident".

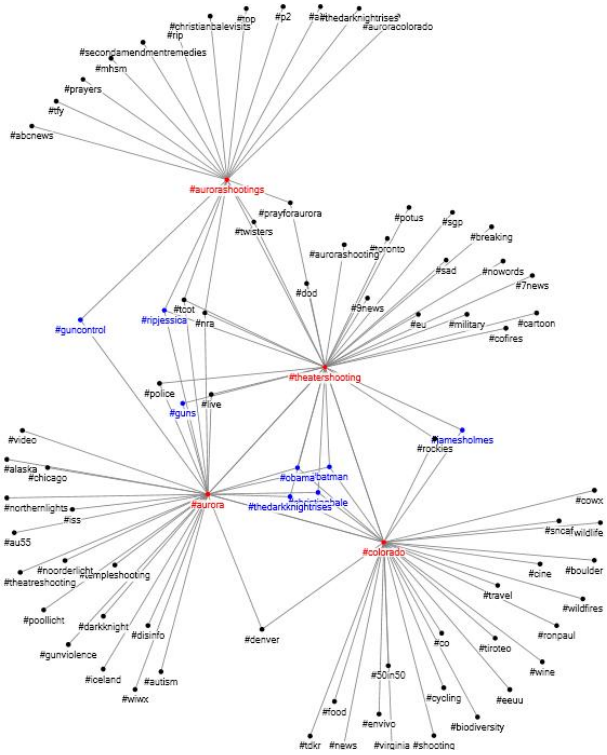

Figure 4. First-level Fruchterman-Reingold graph (case study)

When depicting Table 2, it is rather obvious that the best term related to the "*aurora*" seed is the two-term phrase "*theater shooting*" as derived from the respective hashtag #theatershooting. It was surprising for us that no other search service apart from Reuters suggested this query expansion. It is also worth noticing that phrase "*theater shooting*" was, at the time, suggested by Google after the terms "*aurora colorado*". This is why we consider that suggestions {aurora, theater shooting, colorado} from our algorithm and {aurora, colorado, theater shooting} do match (marked with ✓ in Table 2). We are not strict in suggested term(s)' order of appearance; subject to that the suggested term(s) expand the seed query term. Thus, for the seed "*aurora*" we matched four expansions for Google, three for Yahoo! and Reuters, while only one for Bing and NBC. Reuters query expansion mechanisms surprised us positively once more time, since it was the only service suggested the phrase "*theater shooting*" similarly to our algorithm outcomes. With respect to the seed phrase "*theater shooting*", our algorithm best matched the term sequences {aurora, colorado} (suggested also by Google only) and {aurora, colorado, batman}. The latter is a 3-term query expansion, which no other service provided. It is also worth noticing that our algorithm proposed under a relatively high score (TSW=0.09), a 4-term query expansion {aurora, aurora shootings, colorado, batman}, with

[4] support.google.com/websearch/bin/answer.py?hl=en&answer=106230
[5] NBC search service is powered by Bing

semantically meaningful terms, thus highlighting one major advantage of our proposal. Besides the fact that no other search engine provides a 4-term query expansion (even Google stops after 3 suggested terms), our algorithm expands semantically the user's initial query with terms derived from direct metadata provided by Twitter users. The main benefit derived from above observation is explained due to the fact that we exploit utilization of users' intelligence and capability to describe information, as well as the underlying social intelligence (i.e. through the Twitter social medium) to validate, enhance, or modify it in real-time.As a final validation for the specific case study, we created the graph of cross-related hashtags - as these were provided by the hashonomy - with the help of open-source tool NodeX[6].

Figure 4, illustrates the first-level Fruchterman-Reingold graph, where one may observe that hashtags selected from the initial phase of our algorithm (highlighted in red color) have actually hub properties. The blue highlighted nodes are the hashtags that were candidate terms for query expansion, according to our algorithm and the respective parameters set (k, p, d, TSW_{dist}).

B. Generic evaluation

In order to evaluate our algorithm in a more generic manner, seven human editors were also enrolled in the process. Their task was to subjectively rate the expanded queries in comparison to related terms provided by Google. Each editor was asked to select five different events from Google Hot Searches[7] for the testing period (April 2012). Google Hot Searches displays several top fastest rising searches (and search-terms) by day in the U.S.A.. Each editor had to rate suggested terms/hashtags derived by our algorithm based on a 5-point Likert scale[8] as follows:

1. *Strongly disagree (totally irrelevant suggestion)*
2. *Disagree (not so good suggestion)*
3. *Neither agree nor disagree (nearly same suggestion)*
4. *Agree (potentially better)*
5. *Strongly agree (surely better)*

We ended up with 63 related terms (as provided by Google Hot Searches) in 24 distinct events (11 out of the 35 select events were common). This means that we had an average of 2.63 suggested terms per selected event. We should note here that the rating workload of the editors was not balanced, due to different selected events and thus different amount of related terms. Figure 5 depicts the average evaluator rating for suggested hashtags that presented a TSW score higher or at least equal to 0.25. Considering the evaluation set, we noticed that the majority of subjective rates were close to point 3 in Likert scale (that means "nearly same suggestion" in comparison to Google related terms). However, in many cases the mean evaluator rated the hashtags suggested by our algorithm with a better Likert point scale; this observation means that there were several user-generated terms in hashtag format (single word or concatenated terms) that were more descriptive in comparison to a related term served as query term in Google's log. We have to notice here, that in contrast to the case study experimentation described within the previous sub-section, the related search in the graph was more focused, seeking fewer suggested hashtags in the created semantic networks spanning across a larger time period. Parameter values were set globally for all evaluation cases as follows: k=3 (top-3 results), p=6 (for a-week period) and d=0 (only directly related nodes to the seed).

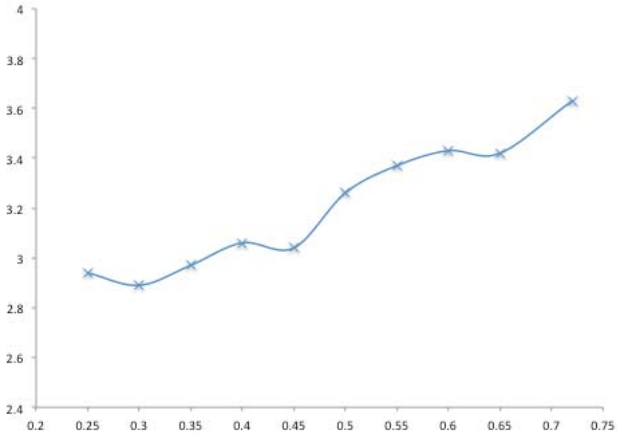

Figure 5. Mean evaluator ratings vs. proposed hashtags ($TSW(k) \geqslant 0.25$)

IV. CONCLUSIONS – FUTURE WORK

In this paper we introduced a query expansion algorithm based on a semantic social network derived by related hashtags generated by Twitter users worldwide. The innovation in this work stands in the fact that we use the users' intelligence and capability to describe information, as well as the social intelligence to validate, enhance, or modify the current information in real-time. We demonstrated basic algorithmic steps involved along with a case study of a tragic incident that was among the breaking news globally for several weeks in 2012. We ended up with a quite promising evaluation, which enrolled human raters and subjective comparisons of suggested results, with respect to related terms provided by Google for similar news events.

Future work includes issues such as an extension towards the semantification of the query expansion mechanism along with intelligent techniques for harnessing crowd wisdom [5], [6], [7]. This practically means, the association of related or synonymous hashtags for their future queries, the hierarchical expression of types and relationships between expanded terms, as well as the involvement of well-known semantic vocabularies like the "friend-of-a-friend" (FOAF) protocol, the Dublin Core Metadata Initiative (DCMI) and others. In parallel with the semantification of our approach, we plan to use the results

[6] http://nodexl.codeplex.com/
[7] http://www.google.com/trends/hottrends
[8] http://en.wikipedia.org/wiki/Likert_scale

derived from the social semantic network for issues like social behavior analysis, social web evolution and trend detection similar to the works described in [8], [9], [10] and [11].

REFERENCES

[1] J. Xu, W. B. Croft, "Query expansion using local and global document analysis", Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96), ACM, New York, NY, USA, 1996, pp. 4-11, doi:10.1145/243199.243202.

[2] R. Jones, B. Rey, O. Madani, W. Greiner, "Generating query substitutions", Proc. 15th International Conference on World Wide Web (WWW '06), ACM, New York, NY, USA, 2006, pp. 387--396, doi:10.1145/1135777.1135835.

[3] R. Baeza-Yates, C. A. Hurtado, M. Mendoza, "Query recommendation using query logs in search engines", Proc. 2004 International conference on Current Trends in Database Technology (EDBT'04), Springer-Verlag Berlin, Heidelberg, 2004, pp. 588-596, doi>10.1007/978-3-540-30192-9_58.

[4] R. Kraft, J. Zien, "Mining anchor text for query refinement", Proc. 13th International Conference on World Wide Web (WWW '04), ACM, New York, NY, USA, 2004, pp. 666-674, doi>10.1145/988672.988763.

[5] M. Hepp, "HyperTwitter: Collaborative Knowledge Engineering via Twitter Messages", Proc. 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW2010), Lisbon, Portugal, 2010, Springer LNCS Vol,. 6317, pp. 451-461.

[6] A. Fuxman, P. Tsaparas, K. Achan, R. Agrawal, "Using the wisdom of the crowds for keyword generation", Proc. 17th International Conference on World Wide Web (WWW '08), ACM, New York, NY, USA, 2008, pp. 61-70, doi:10.1145/1367497.1367506.

[7] M. Pasca, "Organizing and searching the world wide web of facts -- step two: harnessing the wisdom of the crowds", Proc. 16th International Conference on World Wide Web (WWW '07), ACM, NY, USA, 2007, pp. 101-110, doi:10.1145/1242572.1242587.

[8] A. Papantoniou, V. Loumos, M. Poulizos, G. Rigas, "A Framework for Visualizing the Web of Data: Combining DBpedia and Open APIs", Proc. Panhellenic Conference on Informatics 2011, pp. 240-244, http://doi.ieeecomputersociety.org/10.1109/PCI.2011.32.

[9] C.Z. Patrikakis, L. Argyriou and A. Papantoniou, "Online Collaboration", Encyclopedia of Cyber Behavior. IGI Global, 2012, pp. 403-411, doi:10.4018/978-1-4666-0315-8.ch034.

[10] M. Vafopoulos, "Web Science Subject Categorization (WSSC)" Proc. of ACM WebSci, Koblenz, Germany, June 14-17 2011, pp. 1-3.

[11] E. Amarantidis, I. Antoniou, M. Vafopoulos, "Stochastic Modeling of Web evolution", Proc. of 2010 Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA 2010), Chania, Crete, Greece, 8 – 11, June 2010, pp. 45-48.

seed	term1	term2	term3	term4	*TSW (norm.)*	Google	Yahoo!	Bing / NBC	Reuters
aurora									
	theater shooting				0.18	✗	✗	✗	✔
		batman			0.10	✗	✗	✗	✗
		colorado			0.09	✔	✗	✗	✗
			batman		0.08	✗	✗	✗	✗
			aurora shootings		0.07	✔	✗	✗	✗
		darknight			0.09	✗	✗	✗	✗
		9news			0.09	✔	✗	✗	✗
	colorado				0.09	✔	✔	✔	✔
		batman			0.07	✗	✗	✗	✗
		aurora shootings			0.06	✗	✔	✗	✗
	aurora shootings				0.02	✗	✔	✗	✔
		colorado			0.03	✗	✗	✗	✗
			batman		0.04	✗	✗	✗	✗
theater shooting									
	aurora				0.11	✔	✗	✗	✔
		colorado			0.16	✔	✗	✗	✗
			batman		0.15	✗	✗	✗	✗
			aurora shootings		0.13	✗	✗	✗	✗
		aurora shootings			0.08	✗	✗	✗	✗
			colorado		0.08	✗	✗	✗	✗
				batman	0.09	✗	✗	✗	✗
	batman				0.03	✗	✗	✗	✗
	colorado				0.02	✔	✗	✗	✔
		batman			0.08	✗	✗	✗	✗
		aurora shootings			0.05	✗	✗	✗	✗
	darknight				0.01	✗	✗	✗	✗
	9news				0.01	✗	✗	✗	✗
colorado									
	batman				0.65	✗	✗	✗	✗
	aurora shootings				0.35	✔	✔	✗	✗
aurora shootings									
	colorado				0.43	✗	✗	✗	✔
		batman			0.57	✔	✗	✗	✔

Table 2. Query expansion and suggested terms for seeds: #aurora, #theatershooting, #colorado, #aurorashootings (case study)