HASHTAG AND HIGHEST SCORED TERMS FOR EXPANDING QUERY

[1]Wahyu Catur Wibowo & [2]Widodo
[1]Faculty of Computer Science
University of Indonesia, Indonesia
[2]Study Program of Informatics Engineering Education
Universitas Negeri Jakarta, Indonesia

wibowo@cs.ui.ac.id; widodo@unj.ac.id

ABSTRACT

Communicating in short messages, such as using microblogs, was becoming more popular currently. Twitter https://twitter.com supports microblogs and retrieval of the blogs by users. To retrieve Twitter documents, we need specific strategies due to its specific characteristics. One new strategy for improving the effectiveness of twitter document retrieval is using the query expansion technique. This paper elaborates query expansion in twitter document retrieval by using the hashtag. We compared the effectiveness of query expansion in four different scenarios: the baseline result using no query expansion, highest scored-term in terms of frequency-inverse document frequency (*tfidf*), maximum hashtag occurance, and combination of the highest scored-term and the maximum hashtag. The results show that the combination of the maximum term in *tfidf* and the maximum hashtag performs better in retrieving relevant documents than the baseline.

Keywords: Query expansion, maximum hashtag, maximum term.

INTRODUCTION

Microblog, like Twitter, has emerged and become a popular social media to communicate short messages. Each message is known as a tweet. Twitter has some specific characteristics which differs from regular text: time sensitivity, short length, unstructured phrases and abundant information, as described in

Received: 20 January 2016 Accepted: 12 September 2016

(Hu & Liu, 2012). The short length characteristics on Twitter are represented by only allowing a maximum of 140 characters each. This short text often yields less relevance of retrieved documents.

A number of methods have been developed for improving the performance of twitter relevant document retrieval, and researches on query expansion have been conducted for many years for this purpose. There are two major methods in query expansion techniques: global and local methods (Luo et al., 2015). This paper concentrates on using the local methods of query expansion by employing hashtag.

Similar to a microblog, Twitter has some specific features such as hashtag, retweet, mention, URL, and message. These features form the building block of each tweet (Leavitt et al., 2009). A hashtag is the symbol # followed by a keyword that indicates the topic of a tweet. A retweet is a feature to indicate the copying and rebroadcasting of the original tweet, and is preceded by the RT symbol. A mention is a feature which is preceded by the @ symbol to indicate other users as recipients of a tweet. A URL is the url of an address or a website which is embedded in a tweet and is linked to another content. A tweet which contains a hashtag describes a topic that is being stated in the keyword after the hashtag. This definition leads to the intuition that if a query is added with a hashtag keyword, it will increase the relevance of the retrieved documents. In Information Retrieval, the automatic query expansion is utilized to increase the relevant retrieved document. Usually, the method learns from the top-k retrieved document to obtain the term or phrase which will be added to the query.

In the next section, we describe the related works into two sub-sections: related works in microblog retrieval and in query expansion. We focus on our proposed methods: the maximum term in *term frequency inverse-document frequency* (*tfidf*), the maximum hashtag, and the combination of both. We then elaborate our experiment on the proposed methods and evaluate the result using the MAP (Mean Average Precision) metric, and we close the paper with our conclusions based on the result.

RELATED WORKS

We divide this section into two parts: retrieval in microblog, mainly Twitter; and query expansion.

Microblog Retrieval

A microblog, such as Twitter, provides abundant information and gives easy access to users to obtain the information. A user can communicate with other users about certain topics. Usually, a topic is talked over a certain period of time. Based on this intuition, a time sensitive model for retrieving Twitter has been proposed to boost retrieval performance (Shi et al., 2013). They consider that a hot topic can cause an explosion of information in a short time. By using the Bayesian rule as a time factor, it is then considered as a prior probability. The average number of documents delivered in a certain period of time is the key time point to modify the time factor.

To rank retrieved documents, the learning to rank method is proposed in Duan et al. (2010). They assumed that feature selection is an important factor in learning to rank. The three features employed are: content relevant, Twitter specific, and account authority features. Content relevant features are features which describe the relevant content between query and tweets, Twitter specific features are features which describe the specific characteristics of Twitter such as retweet count and url, and account authority features describe the influence of the author of the tweets in Twitter (Leavitt et al., 2009).

Otsuka used one of the specific features of Twitter called hashtag(#) to retrieve information from Twitter (Otsuka et al., 2014). A hashtag is a specific feature of Twitter which indicates a topic. Otsuka proposed a ranking method called Hashtag Frequency – Inverse Hashtag Ubiquity (HF-IHU). Other researches that used specific features of Twitter for retrieving relevant documents are described in Damak et al. (2013); Efron. (2010); Luo et al. (2012); Nagmoti et al. (2010); McCreadie&Macdonald (2013). Damak et al. (2013), used 14 features and their combinations for improving retrieved relevant tweets. Efron (2010) focused on the use of hashtags to assist in query expansion and hashtag association. He reported that the higher the score of hashtag association, the higher the association between this hashtag and others. Luo proposed a building block model of Twitter based on Twitter's specific features (Luo et al., 2012). This building block is called Twitter Building Block (TBB) and is used to rank the retrieved tweets and is successful for increasing the results. Nagmoti proposed a rank based on follower rank, length rank and URL rank (Nagmoti et al., 2010). Follower Rank (FR) indicates the comparison between the sum of followers and the total of followers and the following. Length rank (LR) is a ranking based on the length of a tweet, and URL length is measured by the existence of a URL in a tweet. The combination of these three features yield the best result in relevance retrieval. McCradie focused on one specific feature, the url which is embedded in tweets (McCreadie& Macdonald, 2013). The document relevance with query is increased by linking the tweet to the embedded url. Three methods are proposed for linking to url: Virtual Document, Field-based Weighting and Learning to Rank.

Query Expansion

Query expansion is a method to improve relevant retrieved document by expanding the query input. Basically, a query expansion consists of two methods: global and local (Manning et al., 2008). Researches in this field have long been conducted. Many methods have been developed based on features. Keskustalo (Keskustalo et al., 2015) developed query expansion based on three dimensions: inflectional expansion, historical expansion, and noise expansion. Some researches used external knowledge to enrich document for conducting query expansion (Aggarwal & Buitelaar, 2012); (Qiang et al., 2015); (Weerkamp et al., 2012). External documents such as Wikipedia are used in the research for expanding query (Aggarwal & Buitelaar, 2012); (Weerkamp et al., 2012). Even Weerkamp et al. (2012) used not only Wikipedia, but also web collection, news collection, and blog post collection. Qiang et al. (2015) used freebase as external knowledge. They conducted their research on query expansion in microblog. Research on Twitter for query expansion also has been conducted by many researchers. Some of them focused on Twitter structure (Luo et al., 2012), (Luo et al., 2015). Efronet al. (2010) used hashtag as a Twitter feature for expanding the query. They used hashtags for hashtag query expansion and hashtag association for assisting query expansion.

Our research on query expansion in Twitter, that is not covered in previous researches, is by combining maximum hashtag and best-scored in *tfidf* (term frequency-inverse document frequency).

PROPOSED METHOD

This section discusses our proposed method for expanding query. We discuss the best-n score in *tfidf,* maximum hashtag, and the combination of both.

Best-n **score in*****Tfidf***

Term weighting is an approach to measure the power of a term in a certain category. In information retrieval, term weighting is also used to measure statistically the strength of word appearance in a document (Widodo & Wibowo, 2014). The term weighting approach consists of supervised term weighting and unsupervised term weighting (Xuan & Quang, 2014). The most popular term weighting method is *tfidf* (Lan et all., 2006). This method has two parts: *tf* and *idf*. *Tf* is the frequency of a term appearing in a document which is described in Formula (1). In the formula, f_t indicates the frequency of a term and D_i represents the i^{th} document in which term t is computed. *idf* is Inverse Document Frequency which is the number of documents that contain the investigated term divided by the total number of documents in the collection. Formula (2) explains the *idf* approach. *N* indicates the total number of documents in the corpus and *df* indicates the total number of documents which contain the term *t*. *Tf.idf* is the combination of *tf* and *idf* as shown in Formula (3).

$$tf(t) = f_t(D_i) \tag{1}$$

$$idf(t) = \log(\frac{N}{df}) \tag{2}$$

$$tfidf(t) = f_t(D_i) \times \log(\frac{N}{df}) \tag{3}$$

In information retrieval, the role of *tf* is to improve recall, but not always improving precision. To improve precision *idf* is used because *idf* represents term specificity (Tokunaga &Iwayama, 1994). Hence, in order to improve both, the combination of *tf* and *idf* is employed.

Our method in query expansion was to compute each word in the top 10 documents which were previously retrieved. We added the *best-n* score in *tfidf* to the query as query expansion. We used *n* from 1 to 8 for comparing with the baseline.

Maximum Hashtag

Twitter is a microblog which has several specific features. One of those features is hashtag which is represented by the symbol #. A hashtag is commonly followed by words or phrases without spaces for indicating a topic which is discussed in the tweet. For example, when a tweet contains #big_data, it means that tweet is discussing big data.

Our method for the second query expansion was to compute the maximum appearance of a hashtag in the top 10 documents which were retrieved previously. We called this hashtag a maximum hashtag. This maximum hashtag was added to the query as a second query expansion. We assumed that the most frequent appearance of the hashtag has a high relatedness to the query.

Combination of *tfidf* and Maximum Hashtag

The third query expansion method was to combine the first and second query expansions. The *best-n* in *tfidf* and the maximum hashtag were added to the query. We used again 1 up to 8 best terms in *tfidf* for this approach.

Contribution

We make the following contributions in this paper:

1. We proposed a new method in query expansion in Twitter by combining the maximum hashtag and the highest-scored in *tfidf*.
2. We investigated our method in Bahasa Indonesia, so we know how our method is work in Bahasa Indonesia.
3. We conducted extensive experiments for evaluating our proposed method. The experimental results demonstrate that our proposed method performs better than the baseline.

Methodology

In this section we describe our proposed method, the query expansion based on maximum hashtag and best-n scored term in *tfidf*. Briefly, this query expansion is as shown in Figure 1.

First, we retrieved the tweet documents *T* based on query *q*, while *T* is $\{t_1,t_2,t_3,\ldots,t_n\}$. We ranked this retrieved documents based on cosine similarity. The formula of cosine similarity is shown in Formula (4).

$$sim(T,Q) = \frac{\sum p_i . q_i}{\sqrt{\sum p_i^2 . \sum q_i^2}} \tag{4}$$

Formula *sim(T,Q)* indicates the similarity between query *Q* and tweet *T*, while the value of p_i is obtained from the *tfidf* score of a tweet document and q_i is calculated from the query.

$$T = \{t_1, t_2, t_3, \ldots, t_n\} \tag{5}$$

As shown in Formula (5), *T* is the collection of tweet documents, where t_n is each tweet which is retrieved.

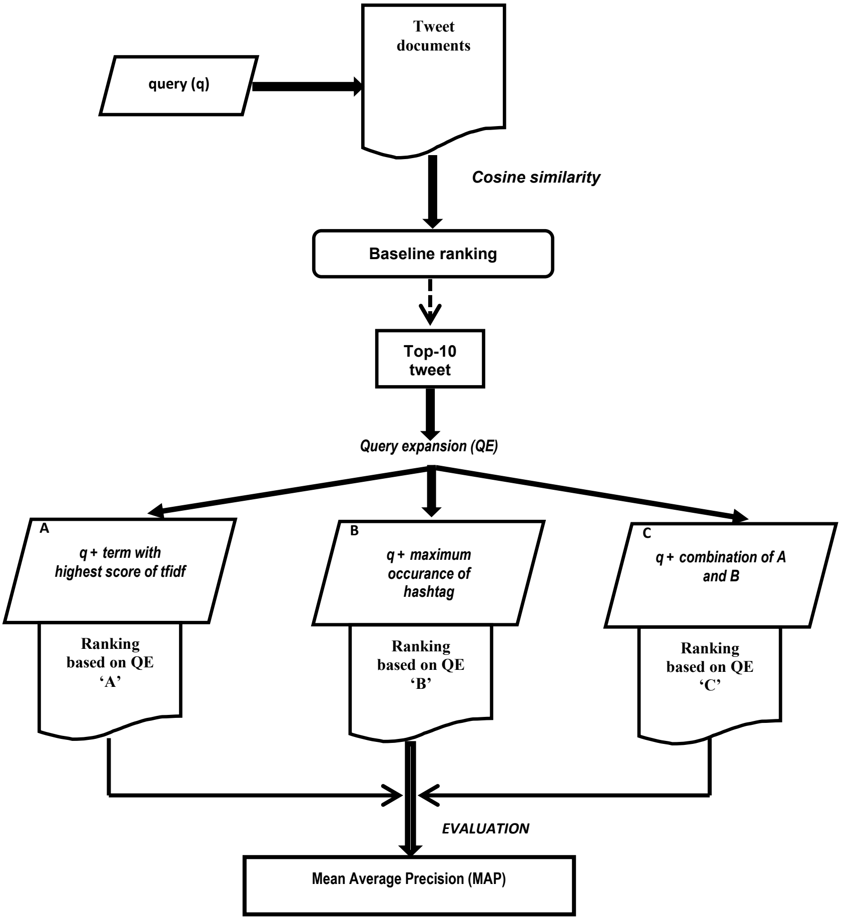

Figure 1. Scheme of proposed method.

Hence we chose the top 10 tweets in the first ranking. We computed the *tfidf* term weighting from this top 10 tweets and we expanded the query q_a by adding the best 1,2, .., until 8 terms to the query.We also computed the frequency of hashtag occurance in the first top 10 tweets of first retrieved. We took this maximum hashtag occurance as query expansion, q_b. Then we combined q_a and q_b as query expansion q_c.

$$q' = q + q_a \quad (6)$$

$$q_a = \max\{score_{tfidf}(T)\} \quad (7)$$

Formulas (6) and (7) show the query expansion as explained previously. q' indicates the query after being added q_a, while q_a is computed based on the best of 3 terms scored by *tfidf*. We assumed that the highest *tfidf* score indicates the terms which have the highest relatedness to the query.

$$q'' = q + q_b \tag{8}$$

$$q_b = \max\{score_H(T)\} \tag{9}$$

Formula (8) indicates the query that is expanded by q_b, while q_b in formula (9) is the most frequent hashtag *H* in the first top 10 tweets. The intuition of this formula is that the hashtag which appears most frequently is the most related to the query.

$$q''' = q + q_c \tag{10}$$

$$q_c = q_a + q_b \tag{11}$$

We also combined q_a and q_b as query expansions as shown in Formulas (10) and (11). From the three query expansion models, the retrieved documents obtained by performing query expansion q', q'', and q''' were ranked. The ranking was evaluated using MAP (Mean Average Precision) and q', q'', and q''' .

EXPERIMENT RESULT

We collected sample data tweet from 25 users and focused on tweets in Bahasa Indonesia. We extracted data from 25 users which contained terms we investigated. For the evaluation method, we used MAP (Mean Average Precision). The higher the value of MAP, the better the performance. We used MAP@10, MAP@20, and MAP@30. In MAP@10 we retrieved the top 10 documents for calculating the value, in MAP@20 we retrieved the top 20 documents, and in MAP@30 we retrieved the top 30 documents. We did not calculate the higher MAP values because it would lead to more non-relevant documents.

The baseline we used was the ranking based on cosine similarity. Then we examined query expansion based on q', q'', and q'''. We used 5 (five) queries in Bahasa Indonesia because these terms or collection of terms were popular terms and key topics in Twitter in Bahasa Indonesia in the period we investigated. We used 5 (five) queries in Bahasa Indonesia as shown in Table 1.

Table 1

Baseline Query

q	Query
q1	Wilayah laut
q2	Pilkada Indonesia hari guru
q3	Peringatan hari guru
q4	Pembajakan kapal
q5	Hepatitis ipb

The result of the baseline is shown in Table 2.

Table 2

MAP of Baseline Query

query	MAP@10	MAP@20	MAP@30
q1	0.91	0.87	0.83
q2	0.85	0.86	0.85
q3	0.79	0.77	0.78
q4	0.64	0.64	0.64
q5	0.95	0.90	0.90

Then we performed our method's hashtag and maximum term query expansion, and the result is as shown in Table 3.

Table 3

Result for Query 1

		Query 1 (q1)							
		1 term	2 terms	3 terms	4 terms	5 terms	6 terms	7 terms	8 terms
q'	**MAP@10**	0.88	0.91	0.91	0.82	0.91	0.95	0.85	0.73
	MAP@20	0.85	0.87	0.86	0.89	0.85	0.87	0.82	0.71
	MAP@30	0.82	0.83	0.82	0.88	0.82	0.83	0.79	0.70
q''	**MAP@10**	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	MAP@20	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
	MAP@30	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83

(continued)

		Query 1 (q1)							
		1 term	2 terms	3 terms	4 terms	5 terms	6 terms	7 terms	8 terms
	MAP@10	0.88	0.91	0.91	0.82	0.91	0.95	0.85	0.73
q’”	**MAP@20**	0.85	0.87	0.87	0.89	0.85	0.87	0.82	0.71
	MAP@30	0.82	0.83	0.83	0.87	0.82	0.83	0.79	0.71

After we had run our method for query 1, the result as shown in Table 3 describes MAP’s score. While *q”* was run, it always yielded the same result in every term added to the query and gave the same MAP’s score with the baseline. However, *q’* and *q”* gave the same result and outperformed the baseline except in 1 term. For MAP@10 and MAP@30, expanded query with 6 terms gave the best result, and for MAP@20, 4 terms were the best result. The next expanded terms tended to decrease. Then, it is obvious that in this query, *q’* and *q’”* have become better methods than *q”*.

Table 4

Result for Query 2

		Query 2 (q2)							
		1 term	2 terms	3 terms	4 terms	5 terms	6 terms	7 terms	8 terms
	MAP@10	1	1	1	0.88	0.95	0.95	0.95	0.95
q’	**MAP@20**	0.99	0.99	0.97	0.83	0.87	0.87	0.88	0.89
	MAP@30	0.96	0.97	0.95	0.84	0.85	0.85	0.86	0.84
	MAP@10	1	1	1	1	1	1	1	1
q”	**MAP@20**	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	MAP@30	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	MAP@10	1	1	1	1	1	1	1	1
q’”	**MAP@20**	1	1	1	0.96	0.96	0.97	1	1
	MAP@30	0.99	0.99	0.99	0.93	0.93	0.94	0.95	0.95

In Table 4, can be obviously seen that *q’*, *q”*, and *q’”* outperform the baseline and *q’”* has the best result. The maximum hashtag (*q”*), like in query 1, has the same results for every expanded term. While *q’”* as the best method in this query gives peak performance at 3 terms expanded, after that expansion the performance tends to be stable.

Table 5

Result for Query 3

		Query 3 (q3)							
		1 term	2 terms	3 terms	4 terms	5 terms	6 terms	7 terms	8 terms
q'	**MAP@10**	1	1	1	1	1	1	1	0.97
	MAP@20	0.98	0.93	0.98	0.98	0.99	0.99	0.99	0.92
	MAP@30	0.96	0.95	0.99	0.94	0.95	0.94	0.95	0.87
q''	**MAP@10**	1	1	1	1	1	1	1	1
	MAP@20	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	MAP@30	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
q'''	**MAP@10**	1	1	1	1	1	1	1	1
	MAP@20	1	1	1	0.99	0.99	0.99	1	1
	MAP@30	0.99	0.99	0.99	0.98	0.98	0.99	0.99	0.97

It is clearly shown in Table 5 that *q''* gives the same result at all level terms just like the previous query. All levels of MAP for *q'*, *q''*, and *q'''* always outperform the baseline and *q'''* has the best performance. It is also shown that that 1 to 3 terms give the best result for *q'''* and 7 terms for *q'*.

Table 6

Result for Query 4

		Query 4 (q4)							
		1 term	2 terms	3 terms	4 terms	5 terms	6 terms	7 terms	8 terms
q'	**MAP@10**	0.54	0.64	0.64	0.77	0.75	0.75	0.75	0.75
	MAP@20	0.56	0.64	0.64	0.76	0.73	0.73	0.73	0.73
	MAP@30	0.56	0.64	0.64	0.76	0.73	0.73	0.73	0.73
q''	**MAP@10**	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79
	MAP@20	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79
	MAP@30	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79
q'''	**MAP@10**	0.69	0.68	0.64	0.85	0.85	0.80	0.79	0.70
	MAP@20	0.69	0.68	0.64	0.80	0.80	0.76	0.73	0.66
	MAP@30	0.69	0.68	0.64	0.80	0.80	0.76	0.73	0.66

The result of query 4 run by our method shown in Table 6 above. The result shows that *q'''* gave the best result and expanded by 4 and 5 terms yielding peak performance while the following expanded terms tended to decrease.

Table 7

Result for Query 5

		Query 5 (q5)							
		1 term	2 terms	3 terms	4 terms	5 terms	6 terms	7 terms	8 terms
	MAP@10	0.91	0.95	0.95	1	1	1	1	1
q'	**MAP@20**	0.87	0.92	0.90	0.96	0.95	0.95	0.96	0.96
	MAP@30	0.87	0.92	0.90	0.96	0.95	0.95	0.96	0.96
	MAP@10	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
q''	**MAP@20**	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
	MAP@30	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
	MAP@10	0.91	0.91	0.95	1	1	1	1	1
q'''	**MAP@20**	0.87	0.86	0.90	0.96	0.955	0.95	0.96	0.96
	MAP@30	0.87	0.86	0.90	0.96	0.955	0.95	0.96	0.96

As shown in Table 7, *q'* and *q'''* give best performance and starting from 3 terms expanded always outperforms the baseline, and 4 terms become the best terms expanded to the query. While more terms expanded to the query, the result of MAP's value tends to be stable.

In our experiments, the combination of the maximum term in *tfidf* and the maximum hashtag (*q'''*) performs better in retrieving relevant documents than the baseline. The number of terms added for expansion to the query to obtain better retrieved relevant documents was four or five.

DISCUSSION

Our experiment results show that performing *q'''* (combination of maximum hashtag and best-scored in *tfidf*) yields the best result in retrieving relevant documents in Twitter. The reason is that maximum hashtag and best-scored in *tfidf* will lead to the nearest relevant documents. The experiment results also show that our method performs best by expanding four or five terms in almost all our experiments. We assumed that expanding more than five terms will decrease the number of relevant documents because the more expanded words given, the more non relevant documents would be extracted. We conducted our methods by processing Twitter in Bahasa Indonesia, so we need to improve our method by comparing it to other languages such as English in the future.

CONCLUSION

Query expansion is a popular method to enhance relevant retrieved documents. Our proposed method in query expansion is to enhance retrieving relevant documents in Twitter. Since the aim was to improve Twitter retrieval, we tried to utilize a feature in Twitter called hashtag. We put the most frequent hashtag in the top 10 previously retrieved documents. The method consisted of query expansion by using the maximum term in tfidf, the maximum hashtag, and a combination of the maximum hashtag and the highest value of *tfidf*. The result of the experiment using five queries in Bahasa Indonesia shows that the combination of the maximum term in *tfidf* and the maximum hashtag (*q'''*) performs better in retrieving relevant documents than the baseline. The number of terms added for expanding the query to obtain better retrieved relevant documents is four or five.

REFERENCES

Aggarwal, N. & Buitelaar, P .(2012). Query expansion using Wikipedia and DBpedia. Retrived from: http://dblp.uni-trier.de/db/conf/clef/clef2012w

Damak, F., Pinel-Sauvagnat, K., & Cabanac, G.(2013). Effectiveness of State-of-the-art Features for Microblog Search”. SAC’13 March 18-22, 2013, Coimbra, Portugal.

Duan, Y, Jiang, L., Qin, T., Zhou, M., & Shum, HY.(2010). An empirical study on learning to rank tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010),* 295–303.

Efron, M. (2010). Hashtag retrieval in a microblogging environment. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieved,*77-78. Switzerland: Geneva.

Hu, X., & Liu, H. (2012). Text analytics in social media. In Aggarwal, C. C., & Zhai, C. X. (Eds.), *Mining Text Data.* doi: 10.1007/978-1-4614-3223-4. Springer.

Keskustalo, H., Kettunen, K., Kumpulainen, S., Ferro, N., Silvello, G., Jarvelin, A., Kekalainen, J., Arvola, P., Saastamoinen, M., & Jarvelin, K. (2015). Targeted query expansions as a method for searching: mixed quality digitized cultural heritage documents. *iConference*.

Lan, M., Tan, C.L., & Low, H.B. (2006). Proposing a new term weighting scheme for text categorization. *Proceedings of the Twenty-First National Conference on Artifical Intelligence (AAAI-06)*, Boston, Massachusetts. pp. 763–768.

Leavitt, A., Burchard, E., Fisher, D., & Gilbert, S. (2009). *The influentials: New approaches for analyzing influence on Twitter*. Retrieved from http://www.webecologicalproject.org

Luo, Z., Osborne, M., Petrovic, S., & Wang, T. (2012). Improving Twitter retrieval by exploiting structural information. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Luo, Z., Yu, Y., Osborn, M., & Wang, T. (2015). Structuring Tweets for improving Twitter search. *Journal of The Association for Information Science and Technology (ASIS&T)*.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. USA: Cambridge University Press.

McCreadie, R. & Macdonald, C. (2013). *Relevance in microblogs: Enhancing Tweet retrieval using hyperlinked documents*. OAIR'2013, 22nd May, 2013*,* Lisbon, Portugal.

Nagmoti, R., Teredesai, A., & De Cock, M. (2010). *Ranking approaches for microblog search.* International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).

Otsuka, E., Wallace, S.A., & Chiu, D. (2014). Design and evaluation of a twitter hashtag recommendation system. *Proceedings of the 18th International Database Engineering & Applications Symposium*, 330-333. Portugal: Porto.

Qiang, R., Fan, F., Lv, C., & Yang, J. (2015). *Knowledge-based query expansion in real-time microblog search*. The Asia Information Retrieval Societies Conference (AIRS).

Shi, C., Xu, B., Lin, H., & Guo, Q. (2013). A time-sensitive model for microblog retrieval. *NLPCC 2013, CCIS 400, pp. 402–409*.

Tokunaga, T., & Iwayama, M. (1994). *Text categorization based on weighted inverse document frequency.* Techical Report 94-TR0001 Department of Computer Science Tokyo Institue of Technology.

Weerkamp, W., Balog, K., & De Rijke, M. (2012). Exploiting external collections for query expansion. *ACM Transactions on the Web*, *6*(4), Article 18. doi: 10.1145/2382616.2.382621

Widodo & Wibowo, W.C. (2014).Improving classification performance by extending documents term. *Proceedings of the International Conference on Data and Software Engineering (ICoDSE)*. ITB, Bandung.

Xuan, N. P., & Quang, H.L. (2014). A new improved term weighting scheme for text categorization. *Knowledge and Systems Engineering 1, Advances In Intelligent Systems and Computing,* 261-270. Springer.