EXPLOITING MULTI-SOURCES QUERY EXPANSION IN MICROBLOGGING FILTERING

Z. Yang[*], *C. Li*[†], *K. Fan*[‡], *J. Huang*[§]

Abstract: Microblogging filtering is intended to filter out irrelevant content, and select useful, new, and timely content from microblogs. However, microblogging filtering suffers from the problem of insufficient samples which renders the probabilistic models unreliable. To mitigate this problem, a novel method is proposed in this study. It is believed that an explicit brief query is only an abstract of the user's information needs, and it's difficult to infer users' actual searching intents and interests. Based on this belief, a filtering model is built where the multi-sources query expansion in microblogging filtering is exploited and expanded query is submitted as user's interest. To manage the external expansion risk, a user filter graph inference method is proposed, which is characterized by combination of external multi-sources information, and a risk minimization filtering model is introduced to achieve the best reasoning through the multi-sources expansion. A series of experiments are conducted to evaluate the effectiveness of proposed framework on an annotated tweets corpus. The results of these experiments show that our method is effective in tweets retrieval as compared with the baseline standards.

Key words: *Microblogging filtering, multi-sources expansion, risk management, matrix factorization*

Received: April 14, 2016 **DOI:** 10.14311/NNW.2017.27.003
Revised and accepted: October 3, 2016

1. Introduction

The widespread use of the Internet has increased the amount of information being stored and accessed through the web in a very fast pace. And the advent of social media (such as Twitter, Weibo, Facebook) has profoundly changed the way people produce and consume information online. The biggest difference between it and

[*]Zhen Yang; College of Computer Science, Beijing University of Technology, Beijing, 100124, China, E-mail: yangzhen@bjut.edu.cn Guangxi Colleges and Universities Key Laboratory of cloud computing and complex systems, Guilin University of Electronic Technology, Guilin 541004, China

[†]Chaoyang Li; College of Computer Science, Beijing University of Technology, Beijing, 100124, China, E-mail: lichaoyang@emails.bjut.edu.cn

[‡]Kefeng Fan – Corresponding author; China Electronics Standardization Institute, Beijing, 100007, China, E-mail: fankf@cesi.cn

[§]Jian Huang; Central University of Finance and Economics, Beijing, 102206, China, E-mail: huangjian0429@cufe.edu.cn

the mainstream news media sites (such as CNN.com or Nytimes.com) lies in that people in social networks are the consumers of information and the producers of information [7,8,17]. This makes the information in the social network disorganized, increasing the difficulty of users to obtain information of interest. And the growing services, such as product review ranking, potential customer targeting and accuracy advertisement pushing in micro-blogging platforms, bring about a growing demand for more effective microblogging filtering technology [34, 40].

Unfortunately, since a microblog usually contains only a limited number of words, the traditional information retrieval (IR) models, including many variants of the vector space model and probabilistic models [24,32], run into difficulty. First, since a microblog is very brief, the vocabulary mismatch problem becomes quite serious . Second, the retrieval model estimation becomes difficult without sufficient word samples. Query expansion is designed to alleviate the words sparseness problem, to obtain a better understanding of the users' intention while avoiding the topic drift caused by the extension. In the past few years, many methods, including the query expansion by the use of a variety of data sources, platforms, and knowledge base, were proposed to improve the performance of microblogging filtering. Zhai C. et al. [43] proposed a family of two-stage language models for information retrieval that explicitly captures the different influences of the query and document collection on the optimal settings of retrieval parameters. Some document expansion methods aim at enriching the evidence of documents. Tao T. et al. [35] proposed a smoothing document language models by analyzing their lexical neighborhoods relying on cosine similarity. Like Tao, Efron M. [12] defined another lexical neighborhood by the likelihood of a document given the language model of each document in the collection. Voorhees E.M. [45], Dalton J. [10], and Gonzalo J. [14] used WordNet as an extension source to extend a query. Navigli R. et al. [21] utilized knowledge ontology to expand query. Bandyopadhyay A. et al. [4] introduced a way of generating query expansion through additional information: a query is submitted to a search engine and the returned results is treated as an extension of the information source to generate query expansion.

Note that external data introduce an expansion risk, i.e., the expanded query could change or drift from the meaning of original query and finally ends up with meanings which are totally different from that of the original one. Thus many researchers focused on how to combine multi-soruces external data to manage the external expansion risk. Wang X. et al. [38] enhanced the effect of negative feedback on the performance of information retrieval in the vector space model and language model. Bendersky M. et al. [6] used multi-source information to improve the quality of retrieval. Weerkamp W. [39] used the news and Wikipedia to extend query together, and achieved good performance. Similar works were also reported as shown in Ref. [25, 29]. However, according to a research on microblogging by Teevan J. [36], the popular search on Twitter is not the most popular topic, but the name of the celebrity. It shows that a large number of users don't search for what interest them by directly using key words. Instead, they are more likely to get the information they need by indirectly following the microblogs of those concerned. This brings a lot of inconvenience to information access. Thus how to provide a timely, effective and accurate method of microblogging filtering is still a open problem.

Though the multi-sources expansion is proved effective, how to use the multi-sources heterogeneous data and how to manage the expansion risk remain open. Motivated by above successful attempts, we propose to tackle this problem by developing an external query expansion method good at risk management. Our contributions are summarized as

- We argue that an explicit brief query is only an abstract of the user's information needs, and it's difficult to infer users' actual searching intents. Instead, we exploit the multi-sources query expansion in microblogging filtering and submit the expanded query as a user's query intentions and then build a corresponding filtering model.

- To manage the external expansion risk, based on the non-negative matrix factorization (NMF) clustering, we propose a user filter graph inference method that combines external multi-sources information, and build a risk minimization filtering model to achieve the best reasoning through the multi-sources expansion.

- We conduct extensive experiments to evaluate the proposed framework on an annotated tweets corpus. With respect to the established baselines, results of these experiments show that our method is effective in the tweets retrieval.

The remainder of this paper is organized as follows. Section 2 gives an overview of the related work. Section 3 introduces the multi-sources query expansion framework based on on-negative matrix factorization (NMF) clustering. To manage the external expansion risk, Section 4 explores the a user filter graph inference method that combins external multi-sources data, and build a risk minimization filtering model to achieve the best reasoning through the multi-sources expansion. Section 5 presents the experimental results. Section 6 is the conclusion.

2. Related work

In this section, we review the milestone achievements in expansion query and relevant query illustration and query expansion milepost related query demonstration.

2.1 Query expansion

Social media information filtering performance can not achieve the expected effect mainly because the search terms input by users can not accurately express their search intention. A well-known breakthrough in this respect is the automatic query expansion (AQE). That is to give users' original query a new description and understanding by adding additional information . This simple method has been successfully applied to the commonly used scheduling model, for example, vector space model [28], probability model [26], statistical language models [42], and random deviation model [1]. AQE has exhibited quite good performance in the field of question and answer, multimedia information retrieval, information filtering, cross language information retrieval etc [31, 44].

2.2 Negative relevant feedback

As extended information has no reliability, feedback and concepts often contain irrelevant information, or ambiguous information. Many scholars expect feedback of irrelevant information and ambiguous information research in order to improve the effect of expansion. Traditional feedback methods such as Rocchio method, already contains the use of negative expansion information. Singhal A. [33] suggested that the expansion process using a non-relevance feedback,can better represent user's query intention than that using only the relevant documentation. As shown in Fig. 1(a),the red dots represent the ideal center of the user's query intent vector while the blue and green dots documentation of negative feedback. The distance between them indicates their similarities. In traditional method, close dots (green dots) is taken positive relevant feedback document while distant dots (blue dots) as negative relevant feedback document. However, we find that query efficiency will not be enhanced with totally irrelevant documentation used as that of negative feedback. Wang X. [38] compared the negative feedback in the language model and the vector space model for their respective effect in improving query performance. As shown in Fig. 1(b), we hope we can, while introducing negative relevant feedback, ignore totally irrelevant feedback represented by the gray dots, for the purpose of approaching user's real query intention through risk management and enhancing query effect. To push further the existing studies, we are expecting to integrate directly and quickly into original query the relevant feedback and negative feedback so that the query performance can be bettered. We will divide feedback document into Positive feedback which is green dots and Negative feedback which is yellow dots and ignore the irrelevant feedback which is gray point, by thinking of risk control and approaching the user's query intention to achieve high retrieval performance effect.

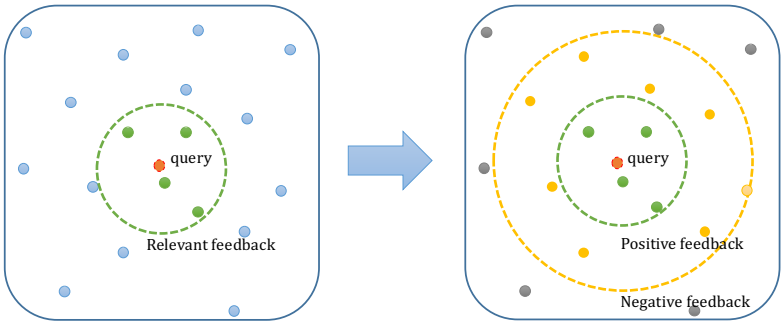

Fig. 1 *Relevant feedback. a) Classic relevant feedback. b) Relevant feedback with positive and negative feedback.*

2.3 Multi-sources query expansion

In the data fusion task [30], the document collections searched by different models should be the same. It is quite different from so-called collection fusion, where the search results come from distinct document collections. Data fusion [46] is important in order to exploit multiple data sources to achieve better performance than any of the individual data source. Several earlier studies have shown significant improvement using data fusion such as on the datasets of TREC-2 and Encyclopedia Britannica [5]. Data fusion can also benefit many information retrieval applications such as meta-search and multimodal search [2, 16].

Different people express the same things differently from different perspectives. Consequently, we can more accurately describe the user query intention through synthesis of these different expressions. As shown in Fig. 2, a user wants to find an apple related tweets and queries may be very short. And we hope to elicit, through external search engine extension, related words such as apple watch, the iPhone 5, and so on. At the same time, when jobs, youtube and other words inevitably appear, we hope we can, through the risk control, limit the introduction of such words, and increase the accuracy of the expansion.

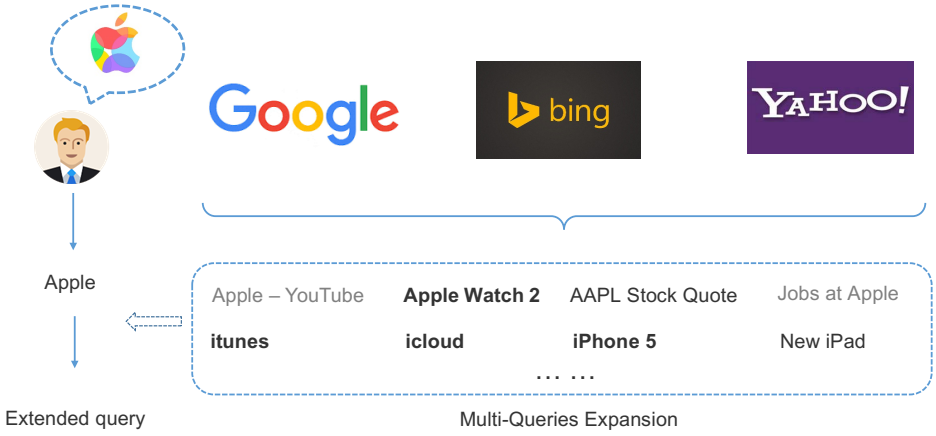

Fig. 2 *An example of multi-sources data based query expansion.*

3. Exploiting multi-sources query expansion in microblogging filtering

In this section, we formally define the problem of microblogging filtering, and discuss how to exploit the multi-sources query expansion in this framework.

3.1 Problem statement

We argue that an explicit brief query which is only an abstract of the user's information needs, and it's difficult to infer users' actual searching intents and interests.

Instead, we exploit the multi-sources query expansion in microblogging filtering and submit the expanded query as a user's interests and then build a corresponding filtering model. In this sense, the general problem we address can be formulated as follows:

Input: Given a specific query, Q, and a candidate microblog set, $\mathbf{C}$, and external multi-sources document sets $\mathbf{E}_i$, which both consists of a set of time stamped microblogs from a variety of social media sources covering a time period. Note $\mathbf{I} \subseteq \mathbf{C}$ is the relevant microblog set selected by query Q.

Output: A ranked version $\mathbf{I}'$ of the stream set $\mathbf{I}$ where each $i_i \in \mathbf{I}'$ is ranked according to its likelihood to give a response to query Q.

3.2 Conventional method

WordNet [20] and Wikipedia [9] are good external knowledge sets and can easily provide a synonym set of words. In recent years, computing word coordinates through word embedding [27] become the new tendency of word similarity computation. The specific training corpus can also provide a set of synonyms. So we take the above three methods as a comparative method and the specific steps are as follows. First, the original query is tagged with Part-of-Speech (POS) labels. Secondly, select a noun in the original query as w, where $w \in Q$, use WordNet to search for the synonym set of w, remove the stop words in the collection and w itself, and construct the extended word set. In the end, the new query retrieval is composed of the original query and the extended word set.

3.3 System framework

To achieve above goal, we build a microblogging filtering system based on multi-sources data query expansion (Fig. 3). This framework has five modules:

- **Step I: Pre-processing**. For input microblogging stream $\mathbf{C}$ and a users' query Q, any non-English symbols, and short texts (length less 2 words) were removed from the text, all words were changed to lowercase and then a simple tokenization method based on white spaces was applied.

- **Step II: Building index**. For input microblogging stream $\mathbf{C}$, we build a index, using Lucene (http://lucene.apache.org), for each microblog and its corresponding id.

- **Step III: Expanding query**. For a users' query Q, we expand original query Q to Q' by the external multi-sources document sets $\mathbf{E}_i$ and the relevant microblog set $\mathbf{I}$.

- **Step IV: Filtering relevant microblog**. Retrieve the relevant microblog set $\mathbf{I}$ by a users' query Q. Goto **Step III** until convergence.

- **Step V: Ranking and evaluation**. Return the ranked version $\mathbf{I}'$ of the stream set $\mathbf{I}$ with by the expanded query Q' and evaluate the filtering performance.

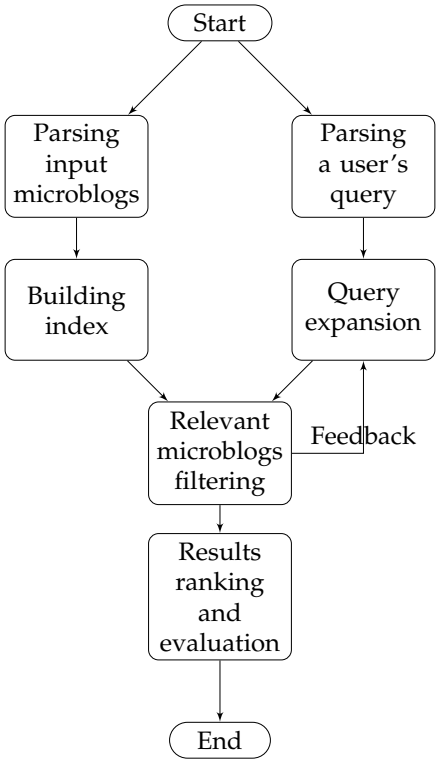

Fig. 3 *A microblogging filtering system based on multi-sources data query expansion.*

Note the query expansion is the key part in the proposed multi-sources query expansion-based microblog filtering framework. We will discuss the details in following section.

4. Multi-sources data based query expansion

To manage the external expansion risk, in this section, we propose a user filter graph inference method that combins external multi-sources data, and build a risk minimization filtering model to achieve the best reasoning through the multi-sources expansion.

As shown in Fig. 4, given a specific query, Q, and a candidate microblog set, $\mathbf{C}$, which consists of a set of time stamped microblogs from a variety of social media sources covering a time period. Note $\mathbf{I} \subseteq \mathbf{C}$ is the relevant microblog set selected by query Q.

Here the goal is to factorize $\mathbf{I}$ into the non-negative $m \times k$ matrix $\mathbf{U}$ and the non-negative $k \times n$ matrix $\mathbf{H}$ that minimize the following objective function:

$$\min_{\mathbf{U},\mathbf{H}} ||\mathbf{I} - \mathbf{U}\mathbf{H}^T||_F^2, \tag{1}$$

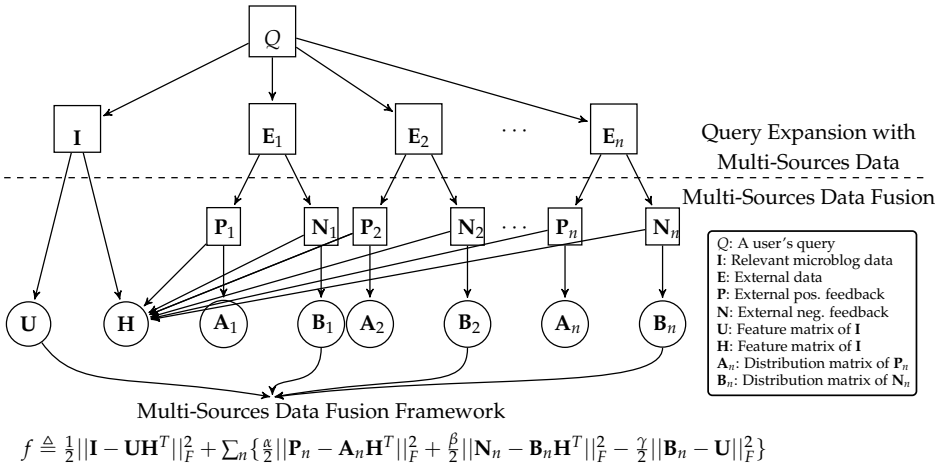

Fig. 4 *Graph inference framework of multi-sources data based query expansion.*

where $||*||_F$ denotes Frobenius norm of a matrix. Note we can use $\mathbf{U}$ as a expanded query.

We argue that an explicit brief query is only an abstract of the user's information needs, and it's difficult to infer users' actual searching intents and interests. We exploit the multi-sources query expansion in microblogging filtering and submit the expanded query as a users' interests and then build a corresponding filtering model. We send a specific query, Q, to the external search engines (Google, Bing, and Yahoo), and obtain the returned search results, i.e., the title and description $\mathbf{E}_n$. For each $\mathbf{E}_n$, we select the top-n pages ($n = 20$) as positive feedback, $\mathbf{P}_n$, and the left as negative feedback, $\mathbf{N}_n$.

Suppose $\mathbf{P}_n$ and $\mathbf{N}_n$ can be factorized on the basis of matrix $\mathbf{H}$,

$$\sum_n ||\mathbf{P}_n - \mathbf{A}_n\mathbf{H}^T||_F^2 + ||\mathbf{N}_n - \mathbf{B}_n\mathbf{H}^T||_F^2. \tag{2}$$

Since $\mathbf{P}_n$ and $\mathbf{N}_n$ are the relevant and non-relevant documents with query Q, thus $\mathbf{B}_n$ should have different distribution with $\mathbf{U}$, we can obtain:

$$\sum_n ||\mathbf{B}_n - \mathbf{U}||_F^2. \tag{3}$$

As shown in Eq. 2 and Eq. 3, with the definition of multi-source query expansion regularization, we propose a microblogging filtering framework. In this sense, the general problem we address can be formulated as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{H}} f = & \frac{1}{2}||\mathbf{I}-\mathbf{U}\mathbf{H}^T||_F^2 \\ & + \sum_n\{\frac{\alpha_n}{2}||\mathbf{P}_n-\mathbf{A}_n\mathbf{H}^T||_F^2+\frac{\beta_n}{2}||\mathbf{N}_n-\mathbf{B}_n\mathbf{H}^T||_F^2-\frac{\gamma_n}{2}||\mathbf{B}_n-\mathbf{U}||_F^2\} \\ s.t. \quad & \mathbf{W} \geq 0, \mathbf{H} \geq 0. \end{aligned} \tag{4}$$

The objective function f can be re-written as follows where the second step of derivation uses the matrix property $Tr(\mathbf{AB}) = Tr(\mathbf{BA})$.

$$f = \frac{1}{2}\{Tr(\mathbf{I} - \mathbf{U}\mathbf{H}^T)(\mathbf{I} - \mathbf{U}\mathbf{H}^T)^T + \sum_n \{\alpha_n Tr(\mathbf{P}_n - \mathbf{A}_n\mathbf{H}^T)(\mathbf{P}_n - \mathbf{A}_n\mathbf{H}^T)^T + \beta_n Tr(\mathbf{N}_n - \mathbf{B}_n\mathbf{H}^T)(\mathbf{N}_n - \mathbf{B}_n\mathbf{H}^T)^T - \gamma_n Tr(\mathbf{B}_n - \mathbf{U})(\mathbf{B}_n - \mathbf{U})^T\}\}. \tag{5}$$

Thus the derivatives of f with respect to $\mathbf{U}$, $\mathbf{H}$, $\mathbf{A}_n$ $\mathbf{B}_n$, are

$$\begin{aligned}
\frac{\partial f}{\partial \mathbf{U}} &= -\mathbf{I}\mathbf{H} + \mathbf{U}\mathbf{H}^T\mathbf{H} + \sum_n (\gamma_i \mathbf{B}_n - \gamma_i \mathbf{U}) \\
\frac{\partial f}{\partial \mathbf{H}} &= -\mathbf{I}^T\mathbf{U} + \mathbf{H}\mathbf{U}^T\mathbf{U} - \sum_n (\alpha_n \mathbf{P}_n^T\mathbf{A}_n + \alpha_n \mathbf{H}\mathbf{A}_n^T\mathbf{A}_n - \beta_n \mathbf{N}^T\mathbf{B}_n + \beta_n \mathbf{H}\mathbf{B}_n^T\mathbf{B}_n) \\
\frac{\partial f}{\partial \mathbf{A}_n} &= -\alpha_n \mathbf{P}_n\mathbf{H} + \alpha_n \mathbf{H}^T\mathbf{H} \\
\frac{\partial f}{\partial B_n} &= -\beta_n \mathbf{N}_n\mathbf{H} + \beta_n \mathbf{B}_n\mathbf{H}^T\mathbf{H} - \gamma_n \mathbf{B}_n + \gamma_n \mathbf{U}.
\end{aligned} \tag{6}$$

Using the Kuhn-Tucker condition $\alpha_{ij}\mathbf{U}_{ij} = 0$ and $\beta_{ij}\mathbf{H}_{ij} = 0$, we get the following equations for $\mathbf{U}_{ij}$, $\mathbf{H}_{ij}$, $\mathbf{A}_{n_{ij}}$, $\mathbf{B}_{n_{ij}}$:

$$\begin{aligned}
&(\mathbf{I}\mathbf{H} + \sum_n \gamma_n \mathbf{U})_{ij}\mathbf{U}_{ij} - (\mathbf{U}\mathbf{H}^T\mathbf{H} + \sum_n \gamma_n \mathbf{B}_n)_{ij}\mathbf{U}_{ij} = 0 \\
&(\mathbf{I}^T\mathbf{U} + \sum_n (\alpha_n \mathbf{P}_n^T\mathbf{A}_n + \beta_n \mathbf{N}_n^T\mathbf{B}_n))_{ij}\mathbf{H}_{ij} - \\
&(\mathbf{H}\mathbf{U}^T\mathbf{U} + \sum_n (\alpha_n \mathbf{H}\mathbf{A}_n^T\mathbf{A}_n + \beta_n \mathbf{H}\mathbf{B}_n^T\mathbf{B}_n))_{ij}\mathbf{H}_{ij} = 0 \\
&(\mathbf{P}_n\mathbf{H})_{ij}\mathbf{A}_{n_{ij}} - (\mathbf{A}_n\mathbf{H}^T\mathbf{H})_{ij}\mathbf{A}_{n_{ij}} = 0 \\
&(\beta_n \mathbf{N}_n\mathbf{H} + \gamma_n \mathbf{B}_n)_{ij}\mathbf{B}_{n_{ij}} - (\beta_n \mathbf{B}_n\mathbf{H}^T\mathbf{H} + \gamma_n \mathbf{U})_{ij}\mathbf{B}_{n_{ij}} = 0.
\end{aligned} \tag{7}$$

These equations lead to the following updating formulas:

$$\begin{aligned}
\mathbf{U}(i,j) &\leftarrow \mathbf{U}(i,j)\sqrt{\frac{(\mathbf{I}\mathbf{H} + \sum_n \gamma_n \mathbf{U})_{ij}}{(\mathbf{U}\mathbf{H}^T\mathbf{H} + \sum_n \gamma_n \mathbf{B}_n)_{ij}}} \\
\mathbf{H}(i,j) &\leftarrow \mathbf{H}(i,j)\sqrt{\frac{(\mathbf{I}\mathbf{U} + \sum_n (\alpha_n \mathbf{P}_n^T\mathbf{A}_n + \beta_n \mathbf{N}_n^T\mathbf{B}_n))_{ij}}{(\mathbf{H}\mathbf{U}^T\mathbf{U} + \sum_n (\alpha_n \mathbf{H}\mathbf{A}_n^T\mathbf{A}_n + \beta_n \mathbf{H}\mathbf{B}_n^T\mathbf{B}_n))_{ij}}} \\
\mathbf{A}_n(i,j) &= \mathbf{A}_n(i,j)\sqrt{\frac{\mathbf{P}_n\mathbf{H}}{\mathbf{A}_n\mathbf{H}^T\mathbf{H}}} \\
\mathbf{B}_n(i,j) &= \mathbf{B}_n(i,j)\sqrt{\frac{\beta_n \mathbf{N}_n\mathbf{H} + \gamma_n \mathbf{B}_n}{\beta_n \mathbf{B}_n\mathbf{H}^T\mathbf{H} + \gamma_n \mathbf{U}}}.
\end{aligned} \tag{8}$$

Thus we can use the NMF general solution to solve the optimization function Eq. (4) and in order to ensure that the decomposition process is positive, we use condition Karush-Kuhn-Tucker (KKT) [41] to get the iterative formula Eq. (8). Note that the KKT condition is the necessary condition for local optimal solutions in non-linear programming, provided that some constraint qualification such as LICQ (linear independence constraint qualification) is satisfied. Furthermore, since $\mathbf{I}$ is non-negative, so $\mathbf{U}$ and $\mathbf{H}$ are non-negative during the updating process. So far, we have proved the correctness of the updating rules in Eq. (8). It can be proven that the updating rules in Eq. (8) are guaranteed to converge. Since the proof process is similar to that in [11], to save space, we omit the detailed proof of the convergence of updating rules in Eq. (8). The details can be seen in Algorithm 1.

Algorithm 1: Multi-sources data based query expansion.

Data: $\mathbf{I}, \mathbf{P}_n, \mathbf{N}_n, \alpha_n, \beta_n, \gamma_n$.
Result: Expanded query $\mathbf{U}$.
begin
  Random initialization $\mathbf{A}_n, \mathbf{B}_n, \mathbf{U}, \mathbf{H}$
  Users' interest description $\mathbf{I}, \mathbf{P}_n, \mathbf{N}_n$
  while *Not convergent* **do**
    Set $\mathbf{T}1 = \mathbf{IH} + \sum_n \gamma_n \mathbf{U}$
    Set $\mathbf{T}2 = \mathbf{UH}^T\mathbf{H} + \sum_n \gamma_n \mathbf{B}_n$
    Set $\mathbf{T}3 = \mathbf{IU} + \sum_n (\alpha_n \mathbf{P}_n^T \mathbf{A}_n + \beta_n \mathbf{N}_n^T \mathbf{B}_n)$
    Set $\mathbf{T}4 = \mathbf{HU}^T\mathbf{U} + \sum_n (\alpha_n \mathbf{H}\mathbf{A}_n^T \mathbf{A}_n + \beta_n \mathbf{H}\mathbf{B}_n^T \mathbf{B}_n)$
    Set $\mathbf{T}5_n = \mathbf{P}_n\mathbf{H}$
    Set $\mathbf{T}6_n = \mathbf{A}_n\mathbf{H}^T\mathbf{H}$
    Set $\mathbf{T}7_n = \beta_n \mathbf{N}_n\mathbf{H} + \gamma_n \mathbf{B}_n$
    Set $\mathbf{T}8_n = \beta_n \mathbf{B}_n\mathbf{H}^T\mathbf{H} + \gamma_n \mathbf{U}$
    Update $\mathbf{U}(i,j) \leftarrow \mathbf{U}(i,j)\sqrt{\frac{\mathbf{T}1(i,j)}{\mathbf{T}2(i,j)}}$
    Update $\mathbf{H}(i,j) \leftarrow \mathbf{H}(i,j)\sqrt{\frac{\mathbf{T}3(i,j)}{\mathbf{T}4(i,j)}}$
    Update $\mathbf{A}_n(i,j) \leftarrow \mathbf{A}_n(i,j)\sqrt{\frac{\mathbf{T}5_n(i,j)}{\mathbf{T}6_n(i,j)}}$
    Update $\mathbf{A}_n(i,j) \leftarrow \mathbf{B}_n(i,j)\sqrt{\frac{\mathbf{T}7_n(i,j)}{\mathbf{T}8_n(i,j)}}$
  return Expanded query $\mathbf{U}$.

5. Experimental settings and results

In this section, we conduct experiments to evaluate the effectiveness of the proposed microblogging filtering framework. Through these experiments, we aim to answer the following questions:

- How effective is the proposed multi-sources data based query expansion framework?

- How does parameter setting affect the performance of the proposed framework?

5.1 Data set

In order to evaluate our methods, we use the a well-known social media data corpus (TREC 2011 Microblog Track Data) [22], that are used in the Microblog Track in TREC 2011 (https://github.com/lintool/twitter-tools/wiki). TREC 2011 microblog track data corpus contains 15989274 cases and 49 test subjects. One example is as follows:

```
<top><num> Number: MB01 </num>
   <title> BBC World Service staff cuts </title>
   <querytime> Tue Feb 08 12:30:27 +0000 2011 </querytime>
   <querytweettime> 34952194402811904 </querytweettime>
   <firstrel>29509222337085440</firstrel>
   <lastrel>34553453812387840</lastrel>
</top>
```

Each of the related tweets are labeled three grades: 0 representative has nothing to do with the topic and 1, 2 represents the relevant and the higher the score the more relevant. With those subjects and corresponding tweets, we therefore can evaluate our method with its variants and other current classic methods.

5.2 Evaluation metrics

We use Map, nDCG and F-measure, which are widely used for information retrieval and filtering tasks, to evaluate the proposed microblogging filtering framework.

The MAP (Mean Average Precision) for a set of queries is the mean of the average precision scores for each query.MAP is computed as follows:

$$\text{MAP} = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}, \tag{9}$$

where Q is the number of queries, and $AveP(q)$ is the average precision at top-100 for each query.

Normalized discounted cumulative gain (nDCG) is a measure of retrieval quality for ranked lists that, in contrast to precision, makes use of graded relevance assessments. nDCG is computed as follows:

$$\text{nDCG} = Z_i \sum_{j=1}^{R} \frac{2^{r(j)} - 1}{\log(1+j)}, \tag{10}$$

where Z_i is a constant to normalize the result to the value of 1. $r(j)$ is an integer representing the relevance level of the result returned at rank j where R is the last possible ranking position. In our experiments, the relevance levels are 0 (irrelevant), 1 (relevant), and 2 (very relevant). nDCG is the primary evaluation metric, which

summarizes precision and ranking. In all experiments we report on nDCG at n calculated using TREC EVAL.

The F-measure is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the fraction of retrieved documents that are relevant to the query. Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P@n, and r is the fraction of the documents that are relevant to the query that are successfully retrieved. The F-measure can be interpreted as a weighted average of the precision and recall, where an F-measure reaches its best value at 1 and worst at 0. Because there are 49 queries in the experiment, we calculate the average F-measure value as the evaluation index.

$$\text{Precision} = \frac{|\{relevant\ documents\} \bigcap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \tag{11}$$

$$\text{Recall} = \frac{|\{relevant\ documents\} \bigcap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \tag{12}$$

$$\text{F} = 2 * \frac{|Precision * Recall|}{|Precision + Recall|} \tag{13}$$

5.3 Performance of multi-sources data based query expansion

In this section, we empirically evaluate the performance of multi-sources data based query expansion method other state-of-art measures. Here we list the methods which we compared our method with and their associated parameter settings.

- **Original query**. For a users' query, any non-English symbols, and short texts (length less 2 words) were removed from the text, all words were changed to lowercase and then a simple tokenization method based on white spaces was applied.
- **Original query+MSDQE**. For a original query, we expand it with multi-sources data based query expansion (MSDQE) method.
- **WordNet query**. For a users' query, we expand it with synsets returned by WordNet [20]. We use WordNet to expand every notional, verbs, adjectives in the query.
- **WordNet query+MSDQE**. For a WordNet query, we expand it with multi-sources data based query expansion (MSDQE) method.
- **Wikipedia query**. For a users' query, we expand it with Wikipedia [13]. Wikipedia is a free-access, free-content Internet encyclopedia, supported and hosted by the non-profit Wikimedia Foundation. We expand the original query into ten related phrases.

- **Wikipedia query+MSDQE**. For a Wikipedia query, we expand it with multi-sources data based query expansion (MSDQE) method.
- **Word2vec query**. For a users' query, we expand it with Word2vec [19]. We use gensim toolkit (https://radimrehurek.com/gensim/models/word2vec.html) to train word2vec model, and thus expand the original query.
- **Word2vec query+MSDQE**. For a Wordvec query, we expand it with multi-sources data based query expansion (MSDQE) method.

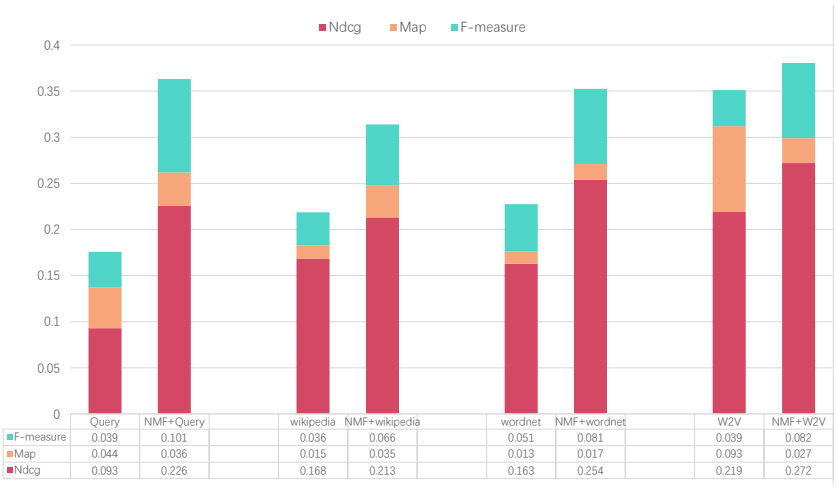

Fig. 5 *Performance of multi-sources data based query expansion.*

The experimental results are shown as stacked bar in Fig. 5 respectively. From this figure, we observe that

- The MSDQE method shows the best nDCG results out of all the methods, i.e., Original query+MSDQE>Word2vec query>Wiki query>WordNet query>Original query.
- The MSDQE method can improve nDCG of all the query expansion methods, i.e., Original query+MSDQE>Original query, Wiki query+MSDQE>Wiki query, WordNet query + MSDQE > WordNet query, and Word2vec query + MSDQE > Word2vec query.

5.4 Parameter setting analysis of the proposed microblogging filtering framework

In this section, we empirically evaluate the effect of parameter setting in the proposed microblogging filtering framework. Since the word2vec query+MSDQE achieved best nDCG performance out of all the methods, we use it as baseline and discuss the effect of parameter setting. Here we list the main parameter settings in the comparsion method used in our experiment.

K	Bing		Google		Yahoo		Map	nDCG	F
	Positive	Negative	Positive	Negative	Positive	Negative			
5	×	×	×	×	×	×	0.166	0.020	0.058
5	✓	×	✓	×	✓	×	0.238	0.023	0.072
5	×	✓	×	✓	×	✓	0.237	0.036	0.069
5	✓	✓	×	×	×	×	0.272	0.02	0.082
5	×	×	✓	✓	×	×	0.273	0.023	0.065
5	×	×	×	×	✓	✓	0.230	0.018	0.075
5	✓	✓	✓	✓	×	×	0.312	0.029	0.067
5	✓	✓	×	×	✓	✓	0.230	0.022	0.066
5	×	×	✓	✓	✓	✓	0.271	0.018	0.075
5	✓	✓	✓	✓	✓	✓	0.242	0.024	0.066
10	✓	✓	×	×	×	×	0.233	0.031	0.070
10	×	×	✓	✓	×	×	0.253	0.029	0.069
10	×	×	×	×	✓	✓	0.242	0.025	0.063
10	✓	✓	✓	✓	×	×	0.245	0.026	0.080
10	✓	✓	×	×	✓	✓	0.232	0.021	0.066
10	×	×	✓	✓	✓	✓	0.251	0.021	0.060
10	✓	✓	✓	✓	✓	✓	0.220	0.026	0.076
15	✓	✓	×	×	×	×	0.262	0.022	0.064
15	×	×	✓	✓	×	×	0.227	0.024	0.075
15	×	×	×	×	✓	✓	0.285	0.024	0.079
15	✓	✓	✓	✓	×	×	0.245	0.022	0.078
15	✓	✓	×	×	✓	✓	0.258	0.027	0.073
15	×	×	✓	✓	✓	✓	0.244	0.020	0.081
15	✓	✓	✓	✓	✓	✓	0.248	0.020	0.071

Tab. I *Microblogging Filtering Performance vs. Parameter Setting.*

- K. We explore how the dimensionality of the low-dimensional representation, K, affect the microblogging filtering probably performance.

- External Data. We send a specific query, Q, to the external search engines (Google, Bing, and Yahoo), and obtain the returned search results, i.e., the title and description E_n. For each E_n, we select the top-n pages ($n = 20$) as positive feedback, P_n, and the left as negative feedback, N_n.

The experimental results are shown as stacked bar in Tab. 5.3 respectively. From this table, we observe that:

- For the dimensionality of the low-dimensional representation, K, in a certain range, the higher the dimensionality of the low-dimensional representation, the better the performance is. In our empirical study, when the dimensionality increase to 10, there is no significant improvement. On the other hand, it is easy to show that the higher the dimensionality, the more computational cost is needed.

- Search engines produce similar performance, and each search engine has its best result in certain applications. Our experiments suggest that every reasonable choice of search engine can work, and results are, insensitive to this choice.

6. Conclusion

In this paper, we propose a framework for query expansion based on multiple sources of external information. We argue that an explicit brief query is only an abstract of the user's information needs, and it's difficult to infer users' actual searching intents and interests. Instead, we exploit the multi-sources query expansion in microblogging filtering and submit the expanded query as a user's interests and then build a corresponding filtering model.

Beside, to manage the external expansion risk, we propose a user filter graph inference method that combines external multi-sources information, and build a risk minimization filtering model to achieve the best reasoning through the multi-sources expansion. We conduct extensive experiments to evaluate the proposed framework on an annotated tweets corpus. With respect to the established baselines, results of these experiments show that our method is effective in the tweets retrieval.

The query expansion framework proposed in this paper is superior in performance, but the algorithm using random matrix as the initial matrix makes the final results fluctuate. In the calculation process algorithm can not automatically adjust parameters and doesn't use parallel calculation method to calculate large amount. We found that query expansion is the key technique of our proposed algorithm. The experiment compared the external knowledge set based query expansion methods such as wiki, WordNet, and based on the frequency distribution of the local expansion method such as Word2Vector. However, since the user query is the focus with the emotion of the user changing [3], our extended content should also change with changes in the user's emotions. The next step, we hope considering the introduction of user behavior [18] and embedding users' implicit interest feature [37] in the query process to help us to further optimize the query results. In addition, in order to optimize the computation efficiency, the algorithm should be implemented on the Hadoop and other big data platform to verify the performance of the algorithm in the production environment [15]. Finally, we hope to introduce the multi-source framework into the field of image, and provide a new light for solve the problem of image segmentation [23].

Acknowledgement

This research was supported by the National Natural Science Foundation of China (No. 61671030), the Excellent Talents Foundation of Beijing, the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (No. CIT&TCD201404052), and the Guangxi Colleges and Universities Key Laboratory of Cloud Computing and Complex Systems (No. 15205).

References

[1] AMATI G., VAN RIJSBERGEN, C.J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*. 2002, 20(4), pp. 357–389, doi: 10.1145/582415.582416.

[2] ASLAM J.A., MONTAGUE M. Models for metasearch. In: D.H. KRAFT, W.B. CROFT, D.J. HARPER, J. ZOBEL, ed. *Proceedings of the 24th International ACM SIGIR Conference*

on Research and Development in Information Retrieval, New Orleans, Louisiana, USA: ACM, 2001, pp. 276–284, doi: 10.1145/383952.384007.

[3] BACH J. A Framework for Emergent Emotions, Based on Motivation and Cognitive Modulators. *International Journal of Synthetic Emotions*. 2012, 3(1), pp. 43–63, doi: 10.4018/jse.2012010104.

[4] BANDYOPADHYAY A., GHOSH K., MAJUMDER P., MITRA M. Query expansion for microblog retrieva. *International Journal of Web Science*. 2012, 1(4): pp. 368–380, doi: 10.1504/IJWS.2012.052535.

[5] BARTELL B.T., COTTRELL G.W., BELEW R.K. Automatic combination of multiple ranked retrieval systems. In: W.B. CROFT, C.J. VAN RIJSBERGEN, ed. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland: ACM, 1994, pp. 173–181, doi: 10.1007/978-1-4471-2099-5_18.

[6] BENDERSKY M., METZLER D., CROFT W.B. Effective query formulation with multiple information sources. *Web Search and Data Mining*. 2012, pp. 443–452, doi: 10.1145/2124295.2124349.

[7] BOSAGH ZADEH R., GOEL A., MUNAGALA K., SHARMA A., et al. On the precision of social and information networks. In: M. MUTHUKRISHNAN, A.E. ABBADI, B. KRISHNAMURTHY, ed. *Proceedings of the first ACM conference on Online Social Networks*, Boston, MA, USA: ACM, 2013, pp. 63–74, doi: 10.1145/2512938.2512955.

[8] CHRISTAKIS N.A., FOWLER J.H. Social network sensors for early detection of contagious outbreaks. *PLOS ONE*. 2010, 5(9), doi: 10.1371/journal.pone.0012948.

[9] COLLINS-THOMPSON K., CALLAN J. Query expansion using random walk models. In: O. HERZOG, H.J. SCHEK, N. FUHR, A. CHOWDHURY, W. TEIKEN, ed. *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*, Bremen, Germany: ACM, 2005, pp. 704–711, doi: 10.1145/1099554.1099727.

[10] DALTON J., DIETZ L., ALLAN J. Entity query feature expansion using knowledge base links. In: S. GEVA, A. TROTMAN, P. BRUZA, C.L.A. CLARKE, K. JÄRVELIN, ed. *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Gold Coast, QLD, Australia: ACM, 2014, pp. 365–374, doi: 10.1145/2600428.2609628.

[11] DING C., LI T., JORDAN M.I. Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding. *Proceedings of the Eighth IEEE International Conference on Data Mining(ICDM'08)*, 2008, pp. 183–192, doi: 10.1109/ICDM.2008.130.

[12] EFRON M., ORGANISCIAK P., FENLON K. Improving retrieval of short texts through document expansion. In: W. HERSH, J. CALLAN, Y. MAAREK, M. SANDERSON, ed. *Proceedings of the 35th International ACM SIGIR conference on Research and development in information retrieval*, Portland, OR, USA: ACM, 2012, pp. 911–920, doi: 10.1145/2348283.2348405.

[13] GABRILOVICH E., MARKOVITCH S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: M.M. VELOSO, ed. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India: AAAI, 2007, pp. 1606–1611, Available from: http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-259.pdf.

[14] GONZALO J., VERDEJO F., CHUGUR I., et al. Indexing with WordNet synsets can improve text retrieval. *Proceedings of the Coling/ACL Workshop on Usage of Wordnet in Natural Language Processing Systems*, 1998, pp. 38–44, Available from: https://arxiv.org/pdf/cmp-lg/9808002.

[15] JAIN A., BHATNAGAR V. Movie Analytics for Effective Recommendation System using Pig with Hadoop. *International Journal of Rough Sets and Data Analysis*. 2016, 3(2), pp. 82–100, doi: 10.4018/IJRSDA.2016040106.

[16] KENNEDY L.S., NATSEV A.P., CHANG S.F. Automatic discovery of query-class-dependent models for multimodal search. In: H. ZHANG, T.S. CHUA, ed. *Proceedings of the 13th Annual ACM International Conference on Multimedia*, Singapore: ACM, 2005, pp. 882–891, doi: 10.1145/1101149.1101339.

[17] KWAK H., LEE C., PARK H., MOON S. What is twitter, a social network or a news media? In: M. RAPPA, P. JONES, ed. *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, NC, USA: ACM, 2010, pp. 591–600, doi: 10.1145/1772690.1772751.

[18] LIU T.Y. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval*. 2009, 8(8), pp. 359–381, doi: 10.1561/1500000016.

[19] MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G.S., DEAN J. Distributed representations of words and phrases and their compositionality. *Neural Information Processing Systems*. 2013, pp. 3111–3119, Available from http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

[20] MILLER G.A. WordNet: a lexical database for English. *Communications of the ACM*. 1995, 38(11), pp. 39–41, doi: 10.1145/219717.219748..

[21] NAVIGLI R., VELARDI P. An analysis of ontology-based query expansion strategies. *Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining*, 2003, pp. 42–49, Available from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.464.8508&rep=rep1&type=pdf#page=44.

[22] OUNIS I., MACDONALD C., LIN J., SOBOROFF I. Overview of the trec-2011 microblog track. *In: Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011, Available from: http://trec.nist.gov/pubs/trec20/papers/MICROBLOG.OVERVIEW.pdf.

[23] PANDA M., HASSANIEN A.E., ABRAHAM A. Hybrid Data Mining Approach for Image Segmentation Based Classification. *International Journal of Rough Sets and Data Analysis*. 2016, 3(2), pp. 65–81, doi: 10.4018/IJRSDA.2016040105.

[24] PONTE J.M., CROFT W.B. A language modeling approach to information retrieval. In: W.B. CROFT, A. MOFFAT, C.J. VAN RIJSBERGEN, R. WILKINSON, J. ZOBEL, ed.*Proceedings of the 21st International ACM SIGIR conference on Research and development in information retrieval*, Melbourne, Australia: ACM, 1998, pp. 275–281, doi: 10.1145/290941.291008.

[25] QI H., LI M., GAO J., LI S. Information retrieval for short documents. *Journal of Electronics (China)*. 2006, 23(6), pp. 933–936, doi: 10.1007/s11767-006-0044-2.

[26] ROBERTSON S.E., VAN RIJSBERGEN C.J., PORTER M.F. Probabilistic models of indexing and searching. In: C.J. VAN RIJSBERGEN, ed. *Proceedings of the 3rd Annual ACM conference on Research and Development in Information Retrieval*, Butterworth & Co. Kent, UK: ACM, 1980, pp. 35–56.

[27] ROY D., PAUL D., MITRA M., GARAIN U. Using Word Embeddings for Automatic Query Expansion. 2016, Available from: https://arxiv.org/abs/1606.07608.

[28] SALTON G., SINGHAL A., MITRA M., BUCKLEY C. Automatic text structuring and summarization. *Information Processing and Management*. 1997, 33(2), pp. 193–207, doi: 10.1016/S0306-4573(96)00062-3.

[29] SCHLAEFER N., CHU-CARROLL J., NYBERG E., et aI. Statistical source expansion for question answering. In: B. BERENDT, A. DE VRIES, W. FAN, C. MACDONALD, I. OUNIS, I. RUTHVEN, ed. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, Glasgow, UK: ACM, 2011, pp. 345–354, doi: 10.1145/2063576.2063632.

[30] SHAW J.A., FOX E.A. Combination of multiple searches. *NIST Special Publication*. 1994, pp. 243–243, Available from: https://www.cs.cmu.edu/afs/cs.cmu.edu/Web/People/nico/pubs/thesis_schlaefer.pdf.

[31] SHEN X., TAN B., ZHAI C. Context-sensitive information retrieval using implicit feedback. In: R.B.-YATES, N. ZIVIANI, ed. *Proceedings of the 28th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, Salvador, Brazil: ACM, 2005, pp. 43–50, doi: 10.1145/1076034.1076045.

[32] SINGHAL A., BUCKLEY C., MITRA M. Pivoted document length normalization. In: H.P. FREI, D. HARMAN, P. SCHAÜBIE, R. WILKINSON, ed. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, Zurich, Switzerland: ACM, 1996, pp. 21–29, doi: 10.1145/243199.243206.

[33] SINGHAL A., MITRA M., BUCKLEY C. Learning routing queries in a query zone. In: N.J. BELKIN, A.D. NARASIMHALU, P. WILLETT, W. HERSH, F. CAN, E. VOORHEES, ed. *Proceedings of the 20th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, Philadelphia, PA, USA: ACM, 1997, pp. 25–32, doi: 10.1145/258525.258530.

[34] SRIRAM B., FUHRY D., DEMIR E., et al. Short text classification in twitter to improve information filtering. In: F. CRESTANI, S. MARCHAND-MAILLET, ed. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, Geneva, Switzerland: ACM, 2010, PP. 841–842, doi: 10.1145/1835449.1835643.

[35] TAO T., WANG X., MEI Q., ZHAI C. Language model information retrieval with document expansion. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, Denver, Colorado, USA, 2006, pp. 407–414, doi: 10.3115/1220835.1220887.

[36] TEEVAN J., RAMAGE D., MORRIS M.R. TwitterSearch: a comparison of microblog search and web search. In: I. KING, ed. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, Kowloon, Hong Kong: ACM, 2011, pp. 35–44, doi: 10.1145/1935826.1935842.

[37] TYAGI S., BHARADWAJ K.K. A Particle Swarm Optimization Approach to Fuzzy Case-based Reasoning in the Framework of Collaborative Filtering.*International Journal of Rough Sets and Data Analysis*. 2014, 1(1), pp. 48–64, doi: 10.4018/ijrsda.2014010104.

[38] WANG X., ZHAI C. Learn from web search logs to organize search results. In: W. KRAAIJ, A.P. DE VRIES, ed. *Proceedings of 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, Netherlands: ACM, 2007, pp. 87–94, doi: 10.1145/1277741.1277759.

[39] WEERKAMP W., DE RIJKE M. Credibility improves topical blog post retrieval. *Meeting of the Association for Computational Linguistics*, 2008, Available from: https://pdfs.semanticscholar.org/f1d5/9980b758d6a2231a786d5202c0c288aca497.pdf.

[40] YANG Z., GAO K., FAN K., LAI Y. Sensational headline identification by normalized cross entropy-based metric. *The Computer Journal*. 2015, 58(4), pp. 644–655, doi: 10.1093/comjnl/bxu107.

[41] YE Y. On the complexity of approximating a KKT point of quadratic programming.*Mathematical Programming and Data*. 1998, 80(2), pp. 195–211, doi: 10.1007/BF01581726.

[42] ZHAI C., LAFFERTY J. A study of smoothing methods for language models applied to information retrieval. *ACM Transaction Information System*. 2004, 22(2), pp. 179–214, doi: 10.1145/984321.984322.

[43] ZHAI C., LAFFERTY J. Two-stage language models for information retrieval. In: K. JÄRVELIN, R. BAEZA-YATES, S.H. MYAENG, ed. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Tampere, Finland: ACM, 2002, pp. 49–56, doi: 10.1145/564376.564387.

[44] ZHAI C., LAFFERTY J. Model-based feedback in the language modeling approach to information retrieval. In: H. PAQUES, L. LIU, D. GROSSMAN, ed. *Proceedings of the Tenth International Conference on Information and Knowledge Management*, Atlanta, GA, USA: ACM, 2001, pp. 403–410, doi: 10.1145/502585.502654.

[45] VOORHEES E.M. Query expansion using lexical-semantic relations. In: W.B. CROFT, C.J. VAN RIJSBERGEN, ed. *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland: ACM, 1994, pp. 61–69, doi: 10.1007/978-1-4471-2099-5_7.

[46] VOORHEES E.M., GUPTA N.K., JOHNSON-LAIRD B. Learning collection fusion strategies. In: E.A. FOX, P. INGWERSEN, R. FIDEL, ed. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, WA, USA: ACM, 1995, pp. 172–179, doi: 10.1145/215206.215357.