BJUT at TREC 2015 Microblog Track: Real-Time Filtering Using Non-negative Matrix Factorization

Chaoyang Li[1,2,3], **Zhen Yang**[1,2,3,*], **Kefeng Fan**[1,4]
1. College of Computer Science, Beijing University of Technology, Beijing 100124, China
2. Beijing Key Laboratory of Trusted Computing, Beijing 100124, China
3. National Engineering Laboratory for CTISCP, Beijing 100124, China
4. China Electronics Standardization Institue, Beijing 100007, China
∗yangzhen@bjut.edu.cn

Abstract

In this paper, we described our approaches to the *Real-Time Filtering Task* in the TREC 2015 Microblog track. We submitted the results for scenario B: periodic e-mail digest. In this ad hoc search task, we introduced a real-time filtering framework using non-negative matrix factorization. To build this framework, we firstly considered the *Real-Time Filtering Task* as a short text retrieval problem, and proposed a non-negative matrix factorization based Microblog retrieval model (NMF Framework). Then after a review of the potential influencing factors in Microblog retrieval, the main influencing factor, i.e., short query expansion, was modeled as the additional regularized constraint items in NMF Framework. Experimental results show the proposed approach is better than classical method in microblog real-time filtering with the above-mentioned additional regularized constraint items.

Introduction

With the rapid growth of micro-blogging users, there are a large number of topics emerging every day. They not only include a small number of hot/stream topics, but also a large number of less popular ones. To help users solve the information overload issue, it is important to recommend personally interesting topics to users.In this sense, this years track is a *Real-Time Filtering Task* under two different situations to explore technologies for monitoring a stream of social media posts with respect to a user's interest profile. We focused on the scenario B: "Periodic email diges" and submitted three results.

Recommendation system is a very mature technology(1; 2; 3).And previous research has studied many methods to solve the problem of hybrid recommendation technique(4; 5; 6).The main idea of our method is to translate the recommendation problem into the Microblog retrieval problem, and the users' interest is considered as the query terms, the retrieval results are considered as the recommended content. First, we construct the initial query through the user interest file, and then combine the different information sources to extend the query. This method uses the accurate descriptions and the ambiguity descriptions of the same user interest with different sources of information to extend the query, alleviates the problem of concept drift in query expansion. In the end, the documents are ranked according to the similarity of the query, and the first one hundred are recommended as the recommended results.

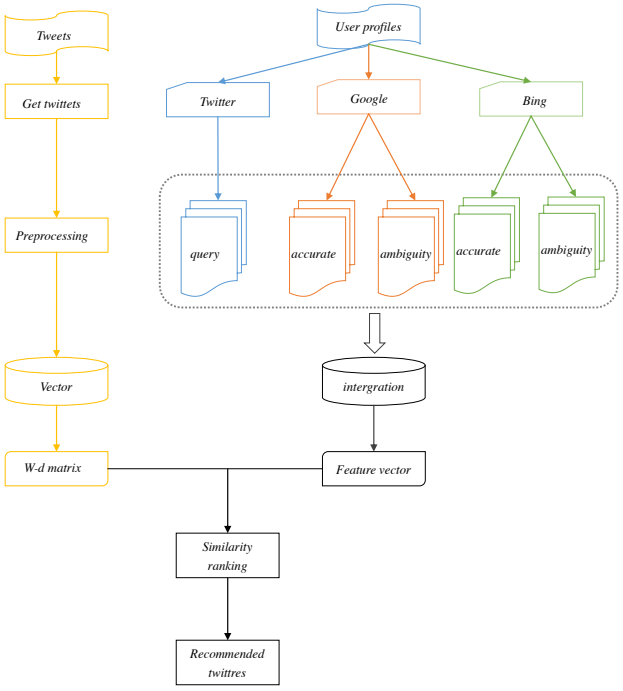

Figure 1: System Framework.

Our Method

System Framework

Figure 1 shows our system framework. It mainly consists of four parts: (1) crawling tweets, (2) pre-processing, (3) analysis of user interest and (4) computing document similarity and ranking.

- Get tweets

  We use the official recommendations to deployment the collect script on EC2 Amazon AWS, through the Twitter's API and obtain real-time microblogging day JSON

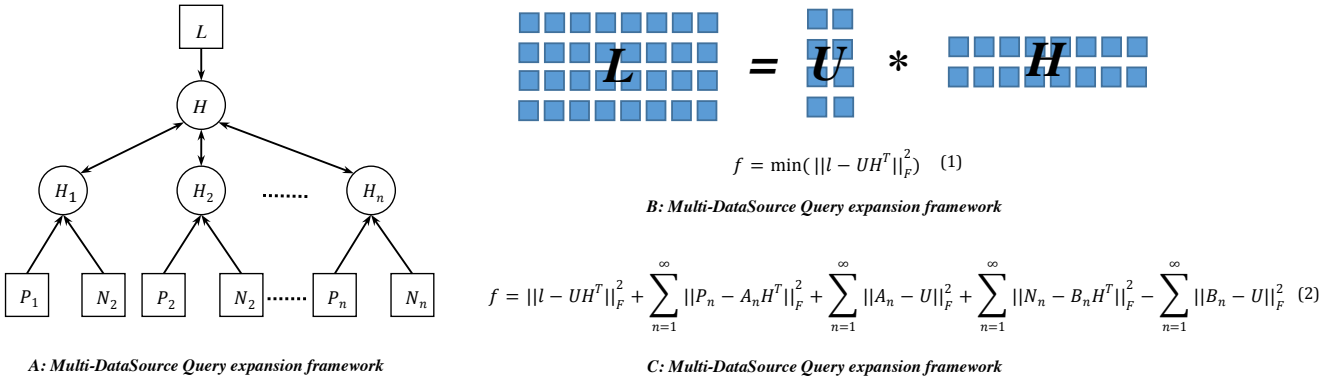

Figure 2: Multi-DataSource Query Expansion.

status flow. After data preprocessing module get the word document matrix and save the stand-by.

- Pre-Processing

  There are four tasks to be done when we deal with one tweet of each topic in our system: (1)removing non English microblogging, (2) removing repeat words over this article 20% microblogging, (3) extracting the tweet id,text and the number of followers, (4) removing the http links,the words of length less than 3 greater than 15 and stopwords, and (5) converting the tweet text to lowercase letters.

- Analysis of user interest

  Interest profiles to be used in the 2015 track will look like traditional TREC topic statements: four fields with the "num" represent the tweet id, "title" containing a few keywords, the "description" containing a one-sentence statement of the information need, and the "narrative", a paragraph-length description of the information need. We are mainly through the following four steps to get the query matrix.

  First of all the interest in the paper is the subject of the word into Google and Bing, the first 100 to get the results of the index as the query expansion document set; secondly,put the interest file in twitter search energy to get back the relevant twetts, the interest in the document and the relevant documents to do the query. If n is less than 20, the 20-n records are added to the query documents. Then, the N $3N$ is used to describe the document set. Finally, the query expansion module is used to get the final query matrix.

- Computing document similarity and ranking

  Calculate each tweet and query matrix cosine similarity, and after the sort, according to the similarity of the blog and the length of the blog post removal of the same tweet. Taking into account the quality of tweet itself, we will have the similarity score, the number of the users followers and the weight of the quality score .After ranking we take the first 100 of the recommended as the results

Multi-Data Source Query Expansion

As we know different search engines express their beliefs over the frame according to their techniques, characteristics, update policies, content preferences, so we search the same world at the different search can get different result lists , We can look at these different result lists as different representations of the query. Based on this we propose a "multi search engine query expansion framewore".

The scheme of our proposed method for extending query for clustering is demonstrated in Figure 2. We will be the result of the search energy Twitter as the original expression of interest, the initial query matrix is denoted as L. As we all know, the search engine will be the exact result of the row in front, so we select the Google and Bing in front of the results as an accurate description of the P, taking the results as a result of the ambiguous description of N. H can be obtained through matrix factorization techniques.

In our application, matrix factorization techniques mapboth terms and short texts to a joint latent factor space of dimensionality K. When ignoring coupling between H, it can be obtained by solving the problem as Formula1 in Figure 2. Where $\| * \|_F$ denotes Frobenius norm of a matrix. Matrices $U \in \mathbf{R}^{m \times k}$ and $H \in \mathbf{R}^{n \times k}$ are the reduced representations for terms and documents respectively in the K dimension joint latent space. Due to NMF the text clustering algorithm,just accepting nonnegative matrices as their inputs, we further add the nonnegative constraints on U and H.

Since the P and the N are different views for the n short texts in the K dimensional latent space, we assume that the different views in the K dimension latent space, A from view P, should be closed to U from the original view L. With this assumption, We can introduce more external data source, which means that N and P can have N, So we sum up the final model as the Formula 2 in Figure 2.

Selection of N

Since the introduction of data source is not only a new feature of query expansion, but also lead to the increase of the number of feature matrix, the N value can not be infinite, the performance should follow the trend of the first increase, due

Table 1: Overall Mean Performances.

	BJUT-nmf1	BJUT-bnmf2
nDCG	0.1008	0.0685

to the time of this experiment is only two groups of extended source, the performance has increased, so the N value is 2.

Similarity distance calculation

In the computation of the final query feature matrix and document matrix similarity, we use the cosine similarity

Submitted Runs and Experiment Results

We submitted two runs: BJUT-nmf1 and BJUT-bnmf2. The only difference between them is that they use a different method of calculating the similarity.BJUT-nmf1 use Cosine similarity and BJUT-nmf1 use sqeuclidean.

Conclusion

In TREC 2015 Real-Time Filtering Track, we submitted two runs.And we apply a NMF Framework which merge multiple source external forward information and negative information into the query.Then Comparison query and results of similarity,Get the final sort.The performances of our two submitted runs are in general better than the median performance. Some of the results are even best results, indicating the effectiveness of our proposed method.

References

[1] Xu J A, Araki K.A SVM-based personal recommendation system for TV programs[C].Multi-Media Modelling Conference Proceedings, 2006 12th International. IEEE, 2006

[2] Wen H, Fang L, Guan L. A hybrid approach for personalized recommendation of news on the Web[J]. Expert Systems with Applications, 2012, 39(5): 5806-5814.

[3] Joachims T. Text categorization with support vector machines: Learning with many relevant features[M]. Springer Berlin Heidelberg, 1998.

[4] Albadvi A, Shahbazi M.A hybrid recommendation technique based on product category attributes[J]. Expert Systems with Applications, 2009, 36(9): 11480-11488.

[5] Sobecki J, Babiak E,Sanina M. Application of hybrid recommendation in web-based cooking assistant[C].Knowledge-Based Intelligent Information and Engineering Systems. Springer Berlin Heidelberg, 2006: 797-804.

[6] Shih Y Y, Liu D R. Hybrid recommendation approaches: collaborative filtering via valuable content information[C].System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on. IEEE, 2005: 217b-217b.