BJUT at TREC 2016: Real-Time Summarization Track

Kai Wang, Zhen Yang*
College of Computer Science, Faculty of Information, Beijing University of Technology, China
yangzhen@bjut.edu.cn

Abstract

This paper describes our approaches to Real-Time Summarization Track in the TREC 2016, including *pushing notifications on a mobile phone task* (Task A) and *periodic email digesting task* (Task B). In Task A, we applied the classifiers to categorize all of the input tweets. External information extracted from Google search engine was well incorporated to enhance the understanding of a users interest. In Task B, all the tweets were classified into a specific topic which was ranked by a scoring system. Finally, we used the non-negative matrix factorization clusering to remove redundancy of the classification results.

Introduction

With the diversification of social media and the popularity of the Internet, data explosion and information overload problems are becoming increasingly serious, so that people cannot quickly and accurately find their own interest in the event-related information. For example, the Web forum because it contains a lot of useful information has became a typical representative of Web. Users are not only users of information in the network forum, but also the creator of information. This makes the network forum a lot of information every day, and information updates fast. A lot of new information makes the forum sensitive to current events, especially some unexpected events. The general users to enter the forum are basically the main post or replies as the object of browsing, the most common form is the same way to browse the forum content. If he wants to understand the whole process of an event and some of the events associated with the description, he may need to browse a large number of topics in order to find the required information, but find out the information may not be complete. The result of this browsing is to spend a lot of time, and not necessarily find satisfactory result. How to effectively summarize from the mass media, such as forums and automated detection of related topics, to help target customers accurate access to the most valuable relevant information abstracts more and more people's attention.

In the field of computer information retrieval, Text Retrieval Conference (TREC) came into being in 1992, along with the development of search technology and the deepening of research, driven by various problems and requirements. This is an activity conducted by the Information Retrieval (IR) community to conduct a search system and user evaluation. It is co-sponsored by the National Institute of Standards and Technology (NIST) and the Advanced Research Projects Agency (DARPA), and is held annually , Its participating groups from many countries, research and academic institutions, government departments and industrial and commercial enterprises, constitute a broad representation of the basis of the search evaluation, the participating units with their own systems for NIST unified corpus and for common tasks To carry out research, and finally by the organizers of NIST to determine the relevance of a unified.

This work is aimed at TREC's Real-Time Summarization Track project and will develop a new system according to the new requirements of the conference. The system can play back the useful, up-to-date and timely update of emergencies in chronological order. Information, so that users can in the event of an emergency, can effectively monitor and event-related information, but also to facilitate public opinion monitoring of government departments, to take preventive measures.

System Framework

It is a real-time job in this years Microblog track that teams listen to the twitter stream via official common API.

- Web Crawler, according to profiles provided by the official crawling the search results as an extended database. For profiles, we extract key words as our basic features.
- Query expansion, Extend the database according to the event, we will deal with the expansion of each event, through the TF-IDF algorithm to query the results of the keywords extracted.
- Text preprocessing, the tweets are processed into the document vectors that we can use.
- Text categorization, According to TREC's official 2015 Twitter data set for classifier training, in the better performance of the classifiers to select three classifiers, as our evaluation classifiers. Each classifier is independent.
- Text clustering, we use the non-negative matrix factorization method to cluster. The relevant tweets in the cluster centers are extracted as our abstracts.

Web Crawler

Short text retrieval suffers severely from vocabulary mismatch problem. Terms overlapping between profiles and tweets are relatively small. Semantic expansion methods can be leveraged to enhance the retrieval performance. Through TREC official title and description of the incident, we have to do an expansion of each event, mainly based on the search engine query results. We take the headline of the incident as a query, the search engine returns the results obtained by the crawler, we only crawl the first N results of the search results, save the documents, and these documents will serve as our event expansion database. Abstract texts are treated as a document, each document contains several terms. After gathering all the documents, we use TFIDF algorithm to calculate TFIDF value of each term for all the profiles. The top k terms of each profiles are used to expand the information.

Query Expansion

Because the data in the tweets are mostly short text, we will encounter a problem in the process of text processing. That is, in the short text, because the text is short, Understanding queries and documents is a loss of semantics, causing the query and the document to lose their original connection. Therefore, we extend the event. Extend the database according to the event, we will deal with the expansion of each event, through the TF-IDF algorithm to query the results of the keywords extracted. Specifically, the TF-IDF value is calculated for each word in the database, and then the first N values are extracted as the result of our query expansion by sorting according to the TF-IDF value.

Text Preprocessing

Due to the large amount of data and the redundancy of the data, the data are preprocessed. First of all, we removed non-English and text length is too short tweet. The filtered text is used to stop the words and stemming, and the processed tweets are used to construct the document word frequency vector according to the word frequency. And we remove non-English and short tweets.

Text Categorization

According to TREC's 2015 tweeter dataset(Zhu et al. 2015), three classifiers were selected as the classifiers. The selected classifiers are: random forests, gradient bo decision trees, and decision trees.

The decision tree is a prediction model; it represents a mapping between object attributes and object values. Each node in the tree represents an object, and each branch path represents a possible attribute value, and each leaf node corresponds to the object represented by the path experienced by the root node to the leaf node value. Decision tree only a single output, if you want a complex output, you can create an independent decision tree to deal with different output. Decision trees in data mining are a common technique that can be used to analyze data and can also be used for forecasting purposes (as the above bank officials use to predict loan risk).

Random forest as the name suggests, is to use a random way to build a forest, there are a lot of forest decision tree composition, and random forest each tree is not associated with the decision tree. After getting the forest, when there is a new input sample, let each decision tree in the forest make a decision to see which class (for the classification algorithm) this sample should belong to, and then see which One class is selected the most, and the sample is predicted for that class.

GBDT (Gradient Boosting Decision Tree), also known as MART (Multiple Additive Regression Tree), is an iterative decision tree algorithm, the algorithm consists of multiple decision trees, and all the tree's conclusions add up to do the final answer. It was proposed to be the beginning of the SVM together with the ability to be considered a generalization algorithm.

Text Clustering

Since we classify each of the tweets with a classifier, the purpose of our text clustering is to deduce the weights of the same or similar meanings in the classified tweets, of tweets. However, because the classification results are not necessarily correct, we hope to use the positive and negative results of the classification to guide our clustering results, so we use non-negative matrix decomposition method to cluster.

The relevant tweets in the cluster centers are extracted as our abstracts.We define the differences between tweets that have different classification results as $C(i,j)$. The conflicts regularization is to minimize the following term as(Tang et al. 2013),

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n} C(i,j)||U(i:) - U(j:)||_2^2 \tag{1}$$

Tweets close to each other in the low-rank space are more likely to have the small the conflict and their distances in the latent space are controlled by their conflicts. For example, $C(i,j)$ controls the latent distance between u_i and u_j. A smaller value of $C(i,j)$ indicates that u_i and u_j are more likely to have the small the conflict according to the property of conflicts. Thus we force their latent representations should be as close as possible, while a larger value of $C(i,j)$ tells that the distance of their latent representations should be larger. After some derivations, we can get the matrix form of conflicts regularization,

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} C(i,j)||U(i:) - U(j:)||_2^2 \\
=&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d} C(i,j)(U(i:) - U(j:))^2 \\
=&\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d} C(i,j)U^2(i,k) \\
&-\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d} C(i,j)U(i,k)U(j,k) \\
=&Tr(U^T LU)
\end{aligned} \tag{2}$$

Where $L = D - Z$ is the Laplacian matrix and D is a diagonal matrix with the i-th diagonal element $D(i,j) = \sum_{j=1}^{n} Z(j,i)$ and $Z(i,j) = \frac{1}{||a[i]-a[j]||}$, a[i] is the label from classifier.

With the definition of conflicts regularization, we propose a framework, based on matrix factorization. The problem is to solve the following optimization,

$$\begin{aligned} \min F = ||A - UV^T||_F^2 + \alpha||U||_F^2 + \beta||V||_F^2 \\ + \gamma Tr(U^T LU) \\ s.t. U \geq 0, V \geq 0 \end{aligned} \tag{3}$$

by removing constants in the objective function,

$$\begin{aligned} F =& Tr(-2A^T UV^T + VU^T UV^T) + \alpha Tr(UU^T) \\ &+ \beta Tr(VV^T) + \gamma Tr(U^T LU) \end{aligned} \tag{4}$$

using the KKT complementary condition and we can get the result of the optimization,

$$AV + \gamma ZD = UV^T V + \alpha U + \gamma DU \tag{5}$$

with define $M = AV + \gamma ZD$ and $N = UV^T V + \alpha U + \gamma DU$ so,

$$U(i,k) \leftarrow U(i,k)\sqrt{\frac{M(i,k)}{N(i,k)}} \tag{6}$$

the same with V,

$$V(i,k) \leftarrow V(i,k)\sqrt{\frac{A^T U}{VU^T U + \beta V}} \tag{7}$$

We alternatively update U and V until achieving convergence. After topic clustering, we select each topic clustering center as the final tweet.

Experiment Results

There are some results of our system. And the evaluation is not complete, the ranking results is as (http://www.trec-ts.org/),

- 20160802 MB226 Q0 760578482128752640 1 246.1 bjutgbdt

- 20160802 MB226 Q0 760269638752149504 2 230.0 bjutgbdt

- 20160802 MB226 Q0 760286323676581889 3 230.0 bjutgbdt

- 20160802 MB226 Q0 760391378438324224 4 230.0 bjutgbdt

Note that our system achieves the score of 246.1, where the 760578482128752640 is the tweetid.

In Task A, we present the results in real time, using classifiers as random forests, decision trees, and gradient boosting decision trees.

Conclusion

In this paper, we presented the implementation details of our runs for Real-Time Summarization Track, the future works emphasis should be on how to improve the accuracy.

References

Tang, J.; Gao, H.; Hu, X.; and Liu, H. 2013. Exploiting homophily effect for trust prediction. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 53–62. ACM.

Zhu, X.; Huang, J.; Zhu, S.; Chen, M.; Zhang, C.; Li, Z.; Dongchuan, H.; Chengliang, Z.; Li, A.; and Jia, Y. 2015. Nudtsna at trec 2015 microblog track: A live retrieval system framework for social network based on semantic expansion and quality model. In *TREC*.