A Time-Enhanced Topic Clustering Approach for News Web Search

Jie Zhao[12], Xiaowen Li[3] and Peiquan Jin[3]

[1]*School of Business, Anhui University*

[2]*School of Management, University of Science and Technology of China*

[3]*School of Computer Science and Technology,*
University of Science and Technology of China

zj_teacher@126.com, xiaowen@mail.ustc.edu.cn, jpq@ustc.edu.cn

Abstract

Time is an important dimension of information space. It plays important roles in Web search, because most Web pages contain time information and many Web queries are time-related. Therefore, exploiting temporal information in Web pages has been a hotspot in the research on Web search. In this paper, we focus on the time-enhanced topic clustering issue for news search results. Traditional clustering algorithms are usually based on the common phrases of Web pages, and they have little consideration about using the temporal information of Web pages. From this perspective, we propose a time-enhanced topic clustering algorithm for news Web pages. It improves traditional algorithms which only consider textual clustering, and applies a temporal clustering procedure on the topics returned by a textual clustering algorithm, which is to arrange every Web page in a cluster along a timeline based on the update time in Web pages. We conduct experiments on a real dataset crawled from Google News, and compare our algorithm with other competitors including K-Means, STC, TFIC, and Minhash Clustering in terms of different metrics such as precision and recall. The experimental results show that the proposed algorithm has better performance under both offline and online clustering test.

Keywords: *Topic clustering, News, Temporal information*

1. Introduction

Web search engines have been mostly used to find information in the Web, in which the search results are usually returned as a list of Web pages. However, as usually a great number of Web pages are returned by a user query, it is very difficult for users to find the appropriate Web pages in the list. Although there have been a lot of ranking algorithms proposed to improve the searching effectiveness, this situation is still becoming worse due to the rapid increasing of Web data volume.

Aiming at solving this problem, researchers proposed to cluster the search results, in which the search results are clustered in terms of several topics, with each topic contains some related Web pages. Traditional topic clustering approaches only consider the textual relevance between query terms and Web pages, and did not consider the time dimension related with Web pages. Since time is one of the essential information dimensions for each Web page, it will be helpful to introduce time into topic clustering and realize time-enhanced topic clustering for search results. For example, given a query term "tennis match", if we use time-enhanced topic clustering approaches, the search results will be further classified and

arranged according to a time line, e.g., “January 2012”, “February 2013”, and so on, on the basis of topic clustering results.

In this paper, we present a time-enhanced topic clustering algorithm for Web search results. In particular, we focus on news Web pages, due to their fast update frequency. The topic of a news Web page is usually focused on some event, people, location, or other entities, and the publish time of a new Web page is very close to or even the same as the event time in the content. As there are often a series of reports regarding a certain event, people usually want to look through the whole development of an event. From this perspective, traditional text-based topic clustering approaches are not suitable for news search, because they only consider the textual relevance and do not take into account the time information related with news Web pages.

The main contributions of the paper can be summarized as follows:

(a) We propose a time-enhanced topic clustering algorithm for news search results, which introduce a further time-based clustering step after conducting a topic clustering for news Web pages. This new algorithm combines textual and temporal clustering techniques, and has better effectiveness in representing news search results.

(b) We conduct experiments on a real dataset crawled from Google News, and compare our algorithm with other competitors including K-Means, STC, TFIC, and Minhash Clustering in terms of different metrics such as precision and recall. The experimental results show that the proposed algorithm has better performance under both offline and online clustering test.

The remaining of the paper is structured as follows. Section 2 introduces the related work. Section 3 presents the time-enhanced topic clustering algorithm. In Section 4, we describe the experimental results, and conclusions and future work are discussed in Section 5.

2. Related Work

Traditional work on Web pages clustering can be classified into four types, namely partition-based clustering, density-based clustering, hierarchy-based clustering, and label-based clustering. Among those algorithms, the typical ones are as follows.

The partition-based clustering algorithm is typically based on the K-Means method [1]. According to those approaches, users have to pre-assign a certain number of partitions, e.g., k. Then we randomly select k documents from the data set, each of them represents the centroid of a partition. Then we compute the distance between each remaining document d and each of the k centroids, and put d into the partition that is nearest to it. However, in this algorithm users must determine the number k at first and the selected centroids have big impact on the performance, therefore its performance is not stable when using different datasets.

In [2], researchers proposed the STC (Suffix Tree Clustering) algorithm. This was first presented in 1999 by Zamir and Etzioni, and they also implemented a prototype called Grouper. The STC algorithm is an online processing one, and is focused on the precision of the description of clusters. It first segments a sentence into a series of sequential words, and chooses those with correct grammar structures and useful meanings as features. Those features (selected sequential words) are put into a suffix tree. The suffix tree is then traversed to get the words that appear in one or more documents more than a given threshold. Those found words are considered as basic clusters, which are further merged into final clusters. The STC algorithm only uses sequential words in a sentence as the base of clustering, which may not correctly reflect the real meaning of the sentence, and in turn worsen the performance of clustering.

Both the above mentioned algorithms did not consider the relationship between separated words in a sentence. From this point of view, the LFIC algorithm (Linguistic Features Indexing Clustering) was proposed [3]. The LFIC algorithm is an improved version on STC. Differing from STC, which looks for the sequential words that appear in one or more documents as basic clusters, LFIC also considers the separated words. It introduced a similarity measurement to determine the importance of the words (either sequential or separated) to a topic. Their experimental results show that LFIC has higher precision than STC.

In 2009, Omar Alonso presented a new visualization way for Web pages clustering [4]. According to this approach, the content time and update time of Web pages are first extracted and the pages are then organized along a time line composed with years, months, and days. This approach provides time-based view for users, but it lacks of the

In 2011, Yahoo! Research presented the SCNSR (Scalable Clustering of News Search Results) [5], a topic clustering system for news. This system involved three stages when performing clustering, i.e., offline clustering, incremental clustering, and online clustering. The experimental results showed that the offline stage had a high recall value but had low precision.

3. Time-Enhanced Topic Clustering

Previous studies such as STC regarded a topic in a news Web page as a series of adjunct words. However, this is not always true. From this point of view, we present a time-enhanced topic clustering algorithm in this paper. Our algorithm is a two-stage solution for news topic clustering, namely an offline stage and an online stage. This is due to that most textual clustering algorithms are executed offline but the clustering of search results requires an online algorithm.

Figure 1 shows the framework of our algorithm. The offline stage is to determine the topics of crawled Web pages, while the online stage is to return time-topic clustered pages.

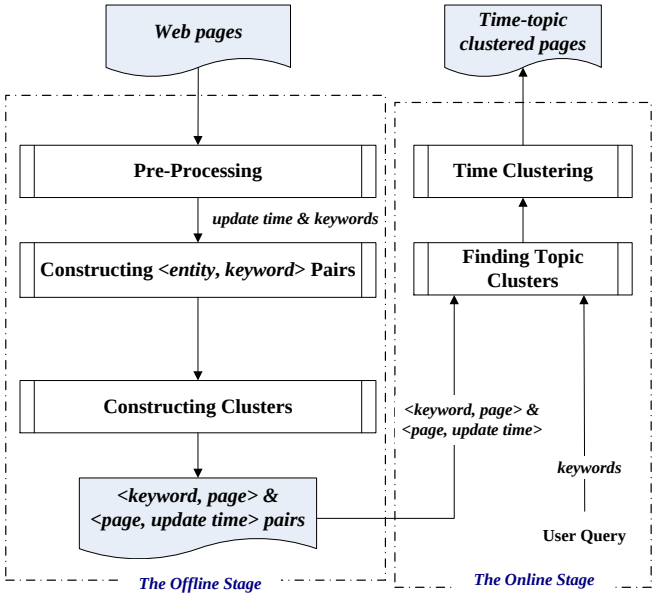

Figure 1. The Framework of Time-enhanced Topic Clustering for News Search Results

As Figure 1 shows, our algorithm consists of three steps:

(1) At the offline stage, we construct a set of <entity, keyword> for each Web page. Each pair of <entity, keyword> represents the relationship between an entity and its related keywords in a Web page. All the <entity, keyword> pairs of the treated Web pages are fused to produce some initial clusters. Those clusters are then merged on the basis of similarity measurement. As a result, we produce a set of clusters for the Web pages.

(2) Next, at the online stage, we refine the initial clusters by introducing the ranking results. After that, we perform a time-based clustering among the Web pages in each cluster.

3.1 Offline Clustering

3.1.1 Constructing Initial Clusters

For each news Web page, we first need to recognize the named entities. This is done by a preprocessing stage. In the preprocessing step, the update time of the Web page is extracted from its tags such as <*meta*>. In order to produce topical information efficiently, we focus on the first paragraph of the Web page. This is because the first paragraph in news report usually includes the main topic of the news.

The first paragraph is segmented into a set of words by using ICTCLAS [6], a very popular tool for Chinese document processing. ICTCLAS can also detect named entities from Web pages. Particularly, we select those words that are labeled in ICTCLAS as *nr*(people), *ns*(location), *nt*(organization), and *nz*(other special names). As a topic is usually described by nouns and verbs, we use nouns (labeled n) and verbs (labeled v) as well as the four types of named entities to describe topics.

The set of Web pages is represented as $P(p_1,p_2,\ldots,p_n)$, where each p_i is a Web page. The named entities extracted from p_i is denoted as $<NE_{i1},NE_{i2},\ldots,NE_{ij}>$, and its keywords are represented as $<keyword_{i1},keyword_{i2},\ldots keyword_{ik}>$. We first filter the keywords that are with low TF-IDF weights. The weight is defined as follows:

$$weight(keyword_{uv}) = tf_{(keyword_{uv})} \times \log \frac{pages_{total}}{pages_u} \quad (u \le n, v \le j) \qquad (3.1)$$

The keywords satisfying $weight(keyword_{mn}) \ge \delta_1$ will be remained in the keywords set, where δ_1 is a predefined threshold. The refined keywords are denoted as $<keyword_{i1}', keyword_{i2}', \ldots, keyword_{im}'>$ ($m \le k$).

After that, we map $<NE_{i1},NE_{i2},\ldots,NE_{ij}>$ with $<keyword_{i1}', keyword_{i2}', \ldots, keyword_{im}'>$ and construct a set of <entity, keyword>, which is denoted as $<I_1,I_2,\ldots,I_N>$. In traditional topic clustering algorithms, when a <entity, keyword> pair, namely I_t, appears in a certain Web page p_i and the appearing times is over a given threshold δ_2, p_i will be associated with I_t. However, they did not consider the textual similarity among different Web pages. So in our approach, we introduce a similarity measurement step on the basis of the traditional process.

3.1.2 Merging the Initial Clusters

In this step, we merge the $<I_1, I_2, \ldots, I_N>$ produced by the preprocessing stage. As there are same entities or keywords among the whole set of <entity, keyword> pairs, we need to find those pairs with common entity or keyword and merge them into one cluster.

The detailed process is as follows. Suppose there are two initial clusters $Cluster_1\{p_1, p_2, \ldots, p_k\}$ and $Cluster_2\{p_1', p_2', \ldots, p_k'\}$, we first compute the centroids of each cluster, e.g., C1 and C2, and then merge them if $Dis\tan ce(C_1, C_2) \le \delta_3$. The distance is determined by the following formula (3.2).

$$Dis\tan ce(C_1, C_2) = \frac{C_1 \cdot C_2}{\|C_1\| \|C_2\|} \tag{3.2}$$

3.1.3 Cluster Refinement

Some Web pages may not be related with a certain topic, even though the entity and keywords describing the topic appear in them. In order to remove such cases, we introduce a cluster refinement procedure, in which the clusters are measured according to the similarity among the Web pages in the cluster.

For a cluster $Cluster_k\{p_{k1}, p_{k2}, \ldots, p_{kn}\}$, suppose p_{ki} $\{keyword_{i1}', keyword_{i2}', \ldots, keyword_{im}'\}$ and p_{kj} $\{keyword_{j1}', keyword_{j2}', \ldots, keyword_{jm}'\}$ are two Web pages in the cluster, we compute the weight of each keyword in those two Web pages in terms of Formula 3.1 and then get the similarity between the two pages using the cosine similarity measurement $Sim_{keyword}(p_{ki}, p_{kj})$. For each Web page p_{ki} in the $Cluster_k$, we compute the similarity between each other Web page in the cluster with p_{ki}.

In addition, we also consider the temporal information when computing the similarity among Web pages. As news Web pages are frequently updated, the Web pages related with the same topic are usually in a certain time period, e.g., in one week. If the update time of a Web page is very close to that of another page, they are very likely to belong in the same cluster. According to this point of view, we define the temporal similarity between two Web pages. Suppose the update time of two Web pages is t_{ki} and t_{kj}, the temporal similarity between the two Web pages is defined as follows:

$$Sim_{temporal}(p_{ki}, p_{kj}) = e^{-|t_{ki} - t_{kj}|/7} \tag{3.3}$$

Then, we define the similarity between two Web pages as the combination of textual similarity and temporal similarity:

$$Sim(p_1, p_2) = Sim_{keyword}(p_{ki}, p_{kj}) \times Sim_{temporal}(p_{ki}, p_{kj}) \tag{3.4}$$

In Formula 3.3, $Sim_{keyword}$ is the textual similarity and $Sim_{temporal}$ is the temporal similarity. When $Sim(p_{ki}, p_{kj}) \ge \delta_3$, we put p_{ki}, p_{kj} into the same cluster. Finally, a cluster $Cluster_k$ will be refined into several new clusters, i.e., $Cluster_k \Rightarrow \{Cluster_{k1}, Cluster_{k2}, \ldots Cluster_{kn}\}$. Those new clusters are further merged into the final clusters at the offline stage.

3.2 Online Clustering

At the offline stage, we partitioned the news Web pages into a set of categories, namely <$Cluster_1$, $Cluster_2$, … , $Cluster_n$>, each of which represents a certain type of event. At the online stage, we conduct further clustering work over the categories. Particularly, when users input keywords and a time range, the search engine will return the results associated with the keywords and time information. Suppose the returned pages are identified as <d_1, d_2, …, d_m>, if $d_k \in Cluster_j$ ($j \le n$, $k \le m$), we put it into the corresponding cluster. As the Web pages in the same category are related with the same topic, and the title of a news page usually can represent the news topic, we simply use the title as the label of the category.

3.3 Time-Based Clustering

During the textual clustering procedure, we get a set of clusters <$Cluster_1$, $Cluster_2$, ... , $Cluster_n$>. In this section, we enhance the clustering effectiveness by adding time-based clustering effects.

According to news Web pages, the update time of the Web pages are usually consistent with the event time reported in their contents. Therefore, in our approach, we use the update time to conduct time-based clustering. At the preprocessing stage, we have already got the update time of the Web pages, namely $UT\{UT_1, UT_2, ..., UT_N\}$, where N is the total count of the Web pages. As a result, we produce a time line for each category $Cluster_i$, which is denoted as $timeline(Cluster_i) = \{\{t\} \mid page \in Cluster_i \& t \in UT_{page}\}$, where t is the time unit. Generally, there are three types of time units, year, month, and day, in the time line. However, as news report is very time sensitive, we only use two small time units, month and day, in our time-based clustering. In particular, we divide the Web pages in $Cluster_i$ according to their update time, and then cluster the pages in each time unit and finally construct a hierarchical view based on the month and day structure.

Based on the time-based clustering, users can find news clusters using textual tags, and browse the pages in each cluster according to a time hierarchical structure. Through this method, users can get to know the evolution history of a news event.

4. Performance Evaluation

4.1 Dataset

We selected 10 news events to form the querying keywords, which described different types of news, such as society, technology, entertainment, finance, and sports. Then we crawled the top 50 results from Google News to produce the dataset. Each crawled page reported one news event. The whole dataset consisted of 500 pages, which are preprocessed to get the update time and named entities using the tool ICTCLAS [6].

The experiment was conducted in Windows 7, with the ObjectWeb Lomboz developing environment and Java support. The tested machine is equipped with an Intel Dual-Core processor and a 2GB memory.

4.2 Results

We measure the performance of offline clustering and online clustering separately. In the offline stage, we use four competitor algorithms:

(1) K-Means, whose count of categories are set to 20.

(2) STC, the term-frequency-based clustering algorithm.

(3) LFIC, the topic-based textual clustering algorithm.

(4) Minhash Clustering, the offline clustering algorithm used in SCNNSR.

We named our algorithm TLFIC. Figure 2 shows the precision of the five algorithms, and Figure 3 shows the recall comparison results.

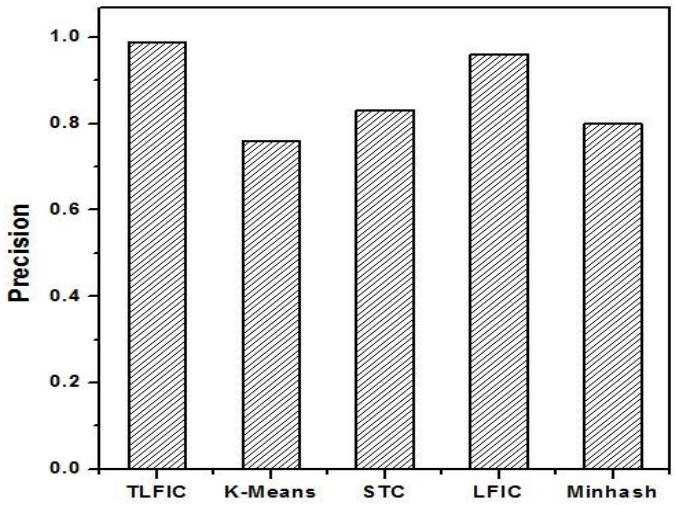

Figure 2. The Precision of the Offline Clustering Algorithms

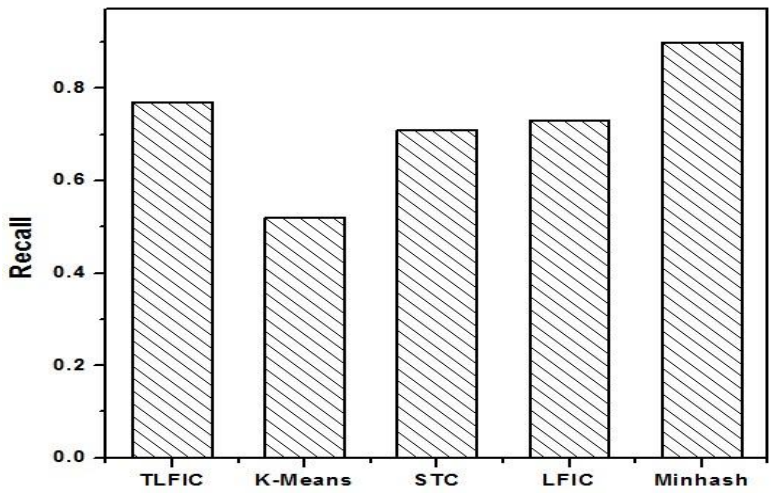

Fig.3 The Recall of the Offline Clustering Algorithms

As shown in Figure 2 and Figure 3, our algorithm has the highest precision. However, the recall of our algorithm is not the highest. This is due to the <entity, keyword> mapping process. As some pages may contain those rarely-appearing entities and keywords, they will not be put into the correct category because of the low similarity between other pages and them.

In the online clustering evaluation, we compare our algorithm with the online part in SCNSR, and the results are shown in Figure 4. The precision and recall of TFLIC is similar to that shown in Figure 2 and Figure 3. Generally, SCNSR needs to construct term vectors for each query and compute the similarity between query and pages, while TFLIC can directly utilize the results of offline clustering and needs not to compute the similarity again, therefore TFLIC can reach a better time performance. Figure 5 shows the snapshot of our time-enhanced clustering algorithm.

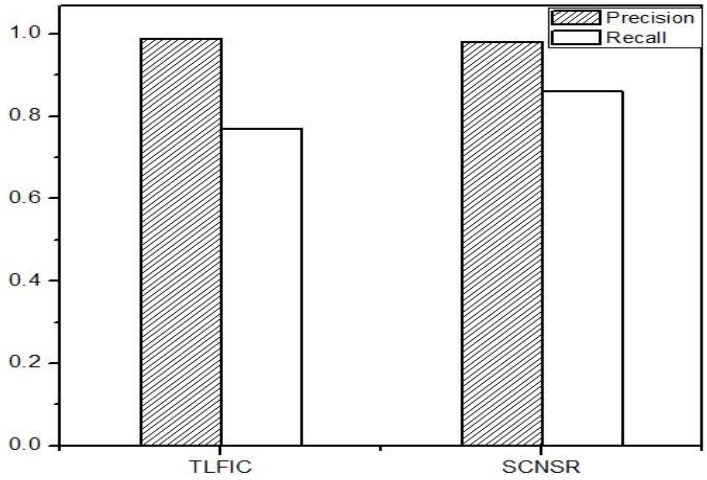

Figure 4. The Results of the Online Clustering Algorithms

Figure 5. The Screenshot of our TFLIC Algorithm

5. Conclusions

Time information can be used in many aspects of Web-related information processing. In this paper, we studied the using of time information in topic clustering of Web pages. We proposed a time-enhanced topic clustering approach to improve the effectiveness of traditional textual clustering ways. Our experimental results showed our approach can reach a better precision compared with other competitor algorithms. Furthermore, our algorithm can return a more structured time line view of Web pages for users, which can better satisfy the users' searching needs.

Acknowledgements

This work is supported by the National Science Foundation of Anhui Province (no. 1208085MG117), the National Science Foundation of China (no. 71273010), the Soft Science Research Program of Anhui Province (grant no. 11020503056), and the USTC Youth Innovation Project.

References

[1] K. Chakrabarti, S. Cauduri and S. won Hwang, "Automatic categorization of query results", In Proceedings of SIGMOD **(2004)**, June 13-18, Paris, France.
[2] O. Zamir and O. Etzioni, "Grouper: A dynamic clustering interface to Web search results", Computer Networks, vol. 31, no. 11–16, **(1999)**, pp. 1361–1374.
[3] S. Zhao, T. Liu and S. Li, "A Topical Document Clustering Method", Journal of Chinese Information Processing, vol. 2, no. 21, **(2007)**.
[4] O. Alonso and M. Gertz, "Clustering and Exploring Search Results using Timeline Constructions", In Proceedings of CIKM **(2009)**, November 2–6, Hong Kong, China.
[5] S. Vadrevu, C. Hui Teo, et al., "Scalable Clustering of News Search Results", In Proceedings of WSDM **(2011)** February 9–12, Hong Kong, China.