Log based Keyword Extraction and Spread based Clustering for an Efficient Information Searching

Nam Hun Park¹ and Kil Hong Joo²

¹, First Author
Dept. of Computer Science, Anyang University, nmhnpark@anyang.ac.kr

², Corresponding Author
Gyeongin National University of Education, khjoo@ginue.ac.kr

Abstract

Today an efficient information search is very important to extract and analyze user requirements in vast amount of web information. Due to this reason, this paper proposes the log based keyword extraction method which finds the associated keywords in a certain domain. Also, this paper proposes the spread based clustering method as clustering the keywords with high association among the keyword-sets of the maximum frequent keyword-sets with redundancy removal by the log based keyword extraction method. The proposed method in this paper is much more efficient and exact than conventional works. The performance of the proposed algorithm is analyzed by a series of experiments to identify its various characteristics.

Keywords: keyword extraction, document clustering, data mining

1. Introduction

With the development of computer and communication technology, the usage of computer and information quantity of users through the network has rapidly increased along. In particular, information surplus with time goes by and efforts to find information needed by a user are being occurred due to explosive increase of electronic information approached through the internet by rapid development of World Wide Web. However, the current information system is not accurately delivering the information needed by the users [1].

The World Wide Web (WWW) not only provides enormous profits but also tremendous challenges for web sites designers and runners. If a particular web site doesn't satisfy the needs of users in a relatively short period of time, the users will quickly move on to another web site [2]. Therefore, it is very important to understand the needs and characteristics of web users [3]. There are many difficulties the search service commonly used today. Because the evaluation of document value by users and knowledge managers on the vast information registered at various fields is not properly implemented. For example, when a user requests for a web search with a certain keyword, a list of more than hundred documents is shown to the user. Also, the important document is not located at the top of the searched list but less important documents are listed at the top not being able to provide desired documents quickly and precisely to the user. Such a problem occurs because the conventional search engine simply prints out the result documents neglecting the relationship (association, distance) such as, 'What role the keyword has, what meaning the keyword has' in the document. Therefore, it makes very difficult for users to find desired information quick and precisely. Thus, it is still challenging to find an efficient searching method on large quantity web text data.
By showing the response list with the keywords in the same classification, the number of response list can be reduced and more precise information can be provided to users. This paper extracts the list of documents approached with certain keywords by analyzing the searching log data generated when approaching the questions for desired documents and response documents. The paper proposes Log based Keyword Extraction Method (LKM) which proposes the searching keywords mutually in close relationship, using Apriori algorithm, the association rule exploring method, after extracting the keyword by performing the morphemes analysis on the documents of the lists, and Spread based Clustering Method (SCM) which finds clusters of keywords with high similarities. This paper is organized as follows.

This paper is organized as follows. Section 2 describes related works. Section 3 introduces the associated keyword extraction and clustering method. In particular, the proposed the log based keyword extraction method and the spread based clustering method are presented in detail. In Section 4, several experiment results are comparatively analyzed to illustrate the various characteristics of the proposed algorithms. Finally, Section 5 draws overall conclusions.

2. Related Works

Cluster analysis is finding a set of objects into classes or clusters based on similarity. Intuitively, objects within a valid cluster are more similar to each other than to an object belonging to a different cluster [4]. The variety of techniques for representing data, measuring similarity between data elements, and grouping data elements have produced a rich and often confusing assortment of clustering methods. It is important to understand the difference between clustering and discriminate analysis [5]. Many researchers have defined four steps for cluster analysis: feature selection or extraction, cluster algorithm design and selection, cluster validation, and result interpretation. These steps are closely related to each other and affect the derived clusters. Several researchers have given significant contribution on the study of cluster techniques [4, 5]. Data mining is a step in Knowledge Discovery in Databases (KDD) process consisting of the application of data analysis and discovery of algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data [6].

Web clustering engines group the result set from the search engine based on their meaning[7]. [8] has a detailed survey of the various clustering engines algorithms. The algorithms are divided into 3 categories namely data-centric, description-aware and description-centric. We have summarized each of these categories here, as they are detailed in [8]. The data-centric algorithms are found to be proven technique targeted for numeric data. Scatter/Gather, Lassi, WebCat, Alsearch are some of the data-centric algorithms. The algorithms typically derives the label for the cluster from its feature vector, which is insufficient from the user perspective, making these algorithms incapable to label and describe the cluster with something sensible for human. The description-aware algorithm concentrates on construction of cluster descriptions which are human interpretable. Suffix Tree Clustering (STC), SnakeT are description-aware algorithms. The setback of the description-aware algorithms is that the clustering precedes and dominates the labeling procedure. The description-centric algorithms are designed to take both the quality of clustering as well description into account. The quality of the description precedes the allocation of the document to the cluster. Lingo, SRC and DisCover are the popular description-centric algorithms. The drawback of these algorithms is that the clusters generated are based on the search results. They may
not correspond to the users’ interest and hence the label may not be very meaningful for the user, as it is generated based on the contents within the cluster. The algorithm does not consider the semantics of the features in order to cluster as well label them.

Current document clustering techniques can be broadly classified into two categories: partitional and hierarchical methods. The popular K-Means [9] and K-Medoid [10] methods are partitional methods which maintain k cluster representatives and assign each document to the cluster with its representative closest to the document. The strong point of the K-Means algorithm is that it is relatively scalable and efficient because it has low time complexity. However, a noise document can substantially influence the generation of a cluster, so it may be difficult to produce a correct result in some cases. The K-Medoid algorithm selects the centers of k clusters initially, and repeatedly replaces one of the selected centers until it finds the best set of k centers. The hierarchical method such as BIRCH [11] confined hierarchy of the clusters in a data set to a tree-like structure. In BIRCH, a CF (Clustering Feature) tree which is used to summarize cluster representations is generated dynamically. After the CF tree is built, any clustering algorithm such as a typical partitioning algorithm is then used. The memory requirement problem of K-Means is resolved and thus the method is suitable for a large data set. However, these algorithms [9-11] are not suitable to solve the document clustering problem since the number of clusters in a set of documents is usually unknown to user [12].

3. Associated Keyword Extraction and Clustering

3.1. Searching log Analysis and Keyword Extraction

Although the amount of searching log files is vast, the number of records with the information of the analysis target according to the purpose is low. Therefore, desired items should be selected and analyzed from the log files according to the analysis purpose. In other words, a data model to explore the association rule necessary for the search should be constructed after removing the unnecessary log items from the overall items of the log files. The items needed here are, ‘User IP address’, ‘Question type’, ‘Searching keyword’, ‘Accessing document number’, and ‘Document access time’. In this paper, the ‘Searching keyword’ item questioned by the user saved in the searching log file and the ensuing documents are defined as 1 transaction. Therefore, a transaction is composed of ‘User IP address’, ‘Searching keyword’, ‘Document access time’ and ‘Accessing document number’, and an appropriate conversion process is required for easy recognition process and analysis of the collected log files as the items are difficultly composed to recognize in the searching log. Figure 1 is an example of an initial searching log data and Table 1 is an example of document accessing details by searching keywords after data conversion.
Table 1. Document Searching Result According to Keywords

<table>
<thead>
<tr>
<th>IP</th>
<th>Keywords</th>
<th>Access time(sec)</th>
<th>#Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>172.16.32.246</td>
<td>Keyword A</td>
<td>27</td>
<td>Document1.doc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Document2.doc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32</td>
<td>Document3.doc</td>
</tr>
<tr>
<td>172.17.4.40</td>
<td>Keyword A</td>
<td>18</td>
<td>Document2.doc</td>
</tr>
<tr>
<td></td>
<td>Keyword B</td>
<td>59</td>
<td>Document5.doc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>Document9.doc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>Document8.doc</td>
</tr>
<tr>
<td>172.17.69.170</td>
<td>Keyword C</td>
<td>22</td>
<td>Document2.doc</td>
</tr>
<tr>
<td></td>
<td>Keyword D</td>
<td>84</td>
<td>Document4.doc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>61</td>
<td>Document3.doc</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

In this paper, time consumed by the user after accessing the document is used as the significant scale to measure the user interesting. The user has searched for more than a certain period of time by accessing the responding document after the user performed the query can be considered as a significant and meaningful document on the concerning query than other documents. Therefore, the refining work of selecting the documents searched for a longer period of time should be performed. If the user has performed query with more than 2 keywords using AND or OR conditions, perform the data correction work and the information cleaning work. Here, perform according to the characteristics of the situation and the data by defining the certain time reference as the Minimum Support.

3.2. Log based Keyword Extraction Method and Spread based Clustering Method

A searching system should be well document-structured for quick and efficient searching. The most important thing to perform document structuralization is the keyword extraction of accurate documents. A document's keyword, as the representing word of the document content, extracting accurate keyword maximizes the efficiency of the searching system[17]. Although the conventional searching engines use the
associated keyword extraction method by applying the association analysis on the keywords by the documents, the result is inaccurate as this does not consider the significance of each document. Therefore, the Log based Keyword extraction Method(LKM) is proposed in this paper in order to overcome such demerits. It can be said the document the user has searched for longer than the minimum support among the searched result list by searching with a certain keyword has a significant meaning to the user or there is high possibility that the document complies with the concerning keyword relative to other documents. Also, all documents are considered as the same class documents, and the method of applying the association analysis to the content of each document is defined as the ‘Document based keyword extraction method(DKM)’.

This paper defines 2 datasets of experiment data and verification data in order to verify the accuracy of clustering. The experiment data implies the raw data to extract the Maximum Frequent Itemset(MFI) and the query log and the overall content of the document is included. Verification data for verification and comparison of the efficiency of the LKM, content collected at different period of time from the experiment data is included. As such, the following variables are defined in this paper in order to calculate the matching rate.

**Definition 1. Matched Count of MFI (MCM)**

The number of MFI in experiment data that complete match with a transaction in verification data defines the matched count of MFI.

**Definition 2. Matched Count of Verification data(MCV)**

The number of transaction in verification data include all keywords of MFI in experiment data defines the matched count of verification data.

The MFI obtained by the experiment data are composed of the keywords and the support of the keywords. The association between these keywords is determined by comparing the keywords from the keywords of the single itemset in the MFI and the keywords from the single transaction in the verification data. In this paper, in order to increase the accuracy of the matching rate with the MFI obtained from the association analysis by the minimum support and the verification data. In addition, this paper proposes the average matched count of MFI as definition 3 and the matched point of MFI as definition 4 to compare the results of LKM and DKM.

**Definition 3. Average Matched Count of MFI(AMCM)**

\[ AMCM = \frac{\sum_{i=1}^{n} MCM_i}{n} (n : \# \text{ of MFI}) \]  

**Definition 4. Matched Point of MFI(MPM)**

\[ MPM = \sum_{i=1}^{n} \frac{100}{MCM_i} (n : \# \text{ of MFI}) \]

\[ Average MPM(AMP) = \frac{MPM}{m} (m : \# \text{ of transaction}) \]

**Definition 5. Itemset Spread Point(ISP)**
From a few to thousands of itemsets exist in the MFI obtained by the LKM, and many similarities exist between the itemsets. In this paper, the Spread based Clustering Method (SCM) is proposed in order to cluster the itemsets with such similarities, and searching efficiency is increased by using this. Figure 2 is the concept diagram which the SCM is illustrated.

The SCM performs clustering based on the similarities, spread and spread score of the 2 itemsets. The similarities imply the level of similarities between the items of a single itemset and those of other itemsets and can be adjusted by the user's characteristics. The spread is defined by the length of the number of the matched count in the verification data, the number of a single itemset is included in the verification data, and the spread score is defined by the total count added to the value obtained by dividing 100 into the verification matched count from each itemset. As the items with the longest spread among the clustering target items are spread most widely with low association between items, the association should be increased first by reducing the spread. Therefore, perform 2nd classification on the longest spread itemsets from the primarily classified itemsets as the clustering targets. Select the itemsets with the highest spread score from the 2nd classification itemsets with same similarities and spread, and cluster the 2 itemsets. As such, repeat the spread based clustering process.
illustrated in Figure 2 until the clustering between the itemsets cannot be performed and ultimately create the clustered itemsets.

4. Experimental Results

The experimental data in this paper uses the two dataset such as the experiment data and the verification data. The experiment data is defined as log and document to extract the maximum frequent itemsets as the LKM proposed in this paper. The verification data is defined as the log and document to verify and compare the efficiency of the LKM. The document searching log and the overall document content during 5 months used to obtain the maximum frequent itemset in the document searching system of D organization actually being operated to obtain more precise experiment result was selected as the experiment data. Also, the data which only the morphemes analysis was applied to the document searching log data of the 5 months generated during the experiment data period and other periods was selected as the verification data in order to evaluate and verify the MFI derived from the experiment data. In order to verify the efficiency of the LKM proposed in this paper, a comparison experiment was performed on the DKM based that performs the association analysis by extracting the keywords within the document in document units.

![Figure 3. MPM and AMPM according to the Minimum Support](image)

The above Figure 3 shows the maximum MPM and AMPM from the MPM by each support obtained by changing the minimum support in 10 seconds interval from 10 seconds to 60 seconds. Although MPM was the highest at the support 20, 40 and 50, as the minimum support 20 seconds was the highest at AMPM. Therefore, the appropriate support in this paper can be 20 seconds. The minimum support was set as 20 seconds in the ensuing experiments.

Figure 4 is the result of AMPM, the MFI matching average count, obtained from the LKM and the DKM based on the verification data. The maximum count of DKM showed 24 in 19% minimum support, and the LKM showed the maximum count of 38 in 25% minimum support. Although the support obtaining the maximum AMPM count was different in each method, it could be known that on the whole, the AMPM counts of the LKM was higher than the counts of the DKM.
Figure 4. Different of AMPM according to the Minimum Support

Figure 5 has calculated AMPM by applying the SCM, clustering with the itemsets with high similarities in the MFI obtained from the LKM proposed in this paper, to the MFI of each minimum support. And then, it compared with AMPM obtained from the MFI of the LKM. The SCM like the result of Figure 5, showed a higher or an identical level of AMPM than the MFI of the LKM in each minimum support. This implies the itemsets clustered by the SCM have a higher association between the items within the itemsets than the itemsets of the associated keywords extracted simply from the association analysis.

5. Conclusion

This paper proposes the log based keyword extraction method as a method of extracting a certain keyword and the associated keyword in a certain domain. Also, this paper propose the spread based clustering method as a method of clustering the items with high association among the itemsets of the maximum frequent itemsets (MFI) with redundancy removed came from the LKM. Currently, the document based method is mostly used for extracting the keyword. Therefore, a comparison experiment was
conducted between the LKM and the DKM in order to verify the effect of the LKM proposed in this paper, and here, the LKM was higher by 42% in the AMPM comparison by the minimum support showing the association of the itemsets. The LKM was higher by 28% in the result which AMPM was compared based on the same number of the MFI, which showed that the LKM had higher association between the keywords than the DKM. It could be known that the method of making into a single transaction by finding a document complying with the purpose of the keyword found by the user searching more than the minimum support was the method of increasing the association between the keywords rather than the method of simply analyzing the association by making into a single transaction by extracting a keyword from a single document, the method of increasing the association between the keywords. As the result of comparing the MFI and AMPM by obtaining the itemsets with the SCM which clusters itemsets with high similarities comparing the similarities between the MFI formed by the LKM, the SCM was shown higher by 12% on average. This shows that similarities exist even for the MFI with redundancy removed, and the items within the itemsets with the similar itemsets clustered increase the association even more. Therefore, the LKM in this paper is much more superior than the association analysis method through the keyword extraction simply within the document.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science Technology(NRF-2011-0025300). In addition, this work was supported by the research fund of Gyeongin National University of Education in 2012.

References

Authors

Nam Hun Park, received the B.S., M.S. and Ph.D. degree in Computer Science from Yonsei University, Seoul, Korea, in 2000, 2002 and 2007. He was a post-Ph.D. at the Department of Computer Science, Worcester Polytech Institute, Worcester, MA. He is currently a professor of Department of Computer Science at Anyang University, Korea. His current interests include mining data streams.

Kil Hong Joo, received the M.S. and Ph.D. degree in Computer Science from Yonsei University, Seoul, Korea, in 2000 and 2004. He is currently a professor of Department of Computer Education at Gyeongin National University of Education, Korea. His current interests include mining data streams, data analysis and smart learning.