Information Retrieval from Text Based Document For Famous Bagan Pagodas in Myanmar

Hnin Wai Zan, Thin Lai Lai Thein
Computer University (Sittwe)
zanzan.gyi@gmail.com,tllthein@gmail.com

Abstract

This paper presents information of most famous Bagan pagodas and the similar information of pagodas among the famous Bagan pagodas by using Apriori Algorithm. Nowadays, people are quite busy as they are occupied with their duties and responsibilities. So they cannot find time to get information. So this system can provide them to get information easily in a very short time. This system will help those who are interested in Bagan pagodas and their background. This paper intends to give information about Bagan pagodas by using Apriori Algorithm .Moreover, this system assists the user knows the founders of pagodas, the type of pagodas, their situation and the time their were constructed . In this system, it will be implemented by using the Association Rule Mining. The system tends to retrieve the information from the many pagodas documents. In this system, it will need the most frequent words to classify the pagodas documents. Most frequent words are got by using the Apriori Algorithm.

Keywords: Association Rule Mining, Apriori Algorithm, Text Mining, Frequent itemsets.

1. Introduction

Information is stored in text databases, which consist of large collections of documents from various sources. Text databases are rapidly growing due to the increasing amount of information available in electronic form. As the amount of text available in electronic form continues to increase at an alarming rate, the tools to manage these textual resources effectively will become critical. Text Mining is necessary to solve that problem.

Text mining is an increasingly important research field because of the necessity of obtaining knowledge from the enormous number of text documents available, especially on the Web. Text mining and data mining, both included in the field of information mining, are similar in some sense, and thus it may seem that data mining techniques may be adapted in a straightforward way to mine text. However, data mining deals with structured data, whereas text presents special characteristics and is unstructured. [3]

The access to a large amount of textual documents becomes more and more effective due to the growth of the Web, digital libraries, technical documentation, medical data, etc. These textual data constitute resources that it is worth exploiting. In this way knowledge discovery from textual databases, or for short, text mining (TM), is an important and difficult challenge, because of the richness and ambiguity of natural language (used in most of the available documents). Therefore, the problem is the existing of huge amount of textual information available in textual form in databases and online sources. [5]

It is not easy to get the information about the historical backgrounds of the pagodas, their situation, the years they were founded , the king where built them in details at one place. In order to solve this problem, this system to retrieve the text based information is implemented. By using this system, the user can know up-to-date information he wants without spending much time.

Association rule mining technique is used to search for interesting relationships among words in a given document collection. Apriori is a classic algorithm for learning association rule mining. In Apriori algorithm frequent subsets are extended one item at a time, and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. [8]

In this paper, the next sections are Section 2 describes related work. Section 3 represents Text Mining .Section 4, proposed system include information retrieval process for the famous Bagan Pagodas using Apriori Algorithm. Section 5 explains system implementation of the system. Section 6, Experiment and result is described. Finally, Section 7 presents the conclusion and references are included.

2. Related Work

Agrawal R and Srikant R. present fast algorithms for mining association rules.[1] Chen, Xin; Wu and Yi-Fang implements personalized knowledge discovery: mining novel association rules from text. [2] In Commercial System, association rule mining is used for mapping one-to-one or many-to-one of the information of the commercial items. To obtain the information of the commercial items by the user desire area or the information of the most popular items in the market. Ismail, Nabil; Mahgoub, Hany; Rösner, Dietmar and Torkey and Fawzy described a text mining technique using association rule extraction.[3] Kusiak and Andrew described association rule the Apriori Algorithm. [4] Nahm, Yong , Un described text mining with information extraction August, 2004.[5] Written, H, Ian presents text mining.[6] Written, H, Ian; Don, J, Katherine; Dewsnip and Michael described text mining in a digital library.[7] In Digital Library System, association rule mining is used for choosing a set of particular topic area in the huge area. Another is to enrich the documents by examining their content, extracting information, and using it to enhance the ways they can be located and presented. Wong, Chung, Pak; Whitney, Paul; Thomas and Jim described visualizing association rules for text mining. [8]

3. Text Mining

Text mining roughly equivalent to text analytics, refers generally to the process of deriving high-quality information from text. Text mining is defined as knowledge discovery in large text collections. It detects interesting patterns such as clusters, associations, deviations, similarities and differences in sets of texts. Text mining concerns the discovery of useful and previously unknown information from unstructured free text and it is strongly related to data mining. [1]

Text mining involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. [6][7]

Association rule mining searches for interesting relationship among data in a given data sets. Association rules are considered interesting if they satisfy both a minimum support threshold (min_sup) and a minimum confidence (min_conf) threshold. Such thresholds can be set by users or domain experts.

It is used to find all frequent itemsets. Each support S of these frequent itemsets will at least equal to a pre-determined min_sup. It will generate the strong association rules from the frequent itemsets. These rules must be the frequent itemsets and must satisfy min_sup and min_conf.

4. Information retrieval process for the famous Bagan Pagodas using Apriori Algorithm

It is not easy to get the information about the historical backgrounds of the pagodas, their situation, the years they were founded , the king where built them in details at one place. In order to solve this problem, this system to retrieve the text based information by using Apriori Algorithm. Moreover, the user can know up-to-date information he wants without spending much time. The following Apriori Algorithm is applied in this system.

Begin
Tokenized words in famous Bagan Pagodas information;
Computing frequent wordsets in given information;
Begin
L_k = null; k = 0;
C_1 includes all information of pagodas;
L_1 includes words greater than 1 in C_1;
While L_{k+1} is not empty
C_{k+1} includes (first word, second word);
L_{k+1} includes wordsets greater than 1 in C_{k+1};
K^{++};
Return same information of pagodas;
End;
End.

So, the user can know the history of Bagan pagodas within a short period.

5. Implementation of the System

This paper implements the Association Rule Mining within a system that classifies the Bagan Pagodas. In this implementation use the Apriori Algorithm to search the main frequent words that are used to classify.

In this paper;

The system shows the pagoda's list to the user to choice. The user choice the pagodas that the user desires to classify. Then the system will show the documents of the pagoda's information. After

choice, the system will be prepared to classify the pagodas. The system will tokenize each pagoda's document to words and remove the unnecessary stop words. Then the system will choose the main data (words) and remove the not important words. After that, the system stores that word in the Pagoda table in the Pagoda database for each document. Then, use the Apriori Algorithm to get the most frequent word that will use to classify. After that, the system will store that frequent words in the WordCount table. Then the system will compare the frequent words from the WordCount table with the words from Pagoda table. If each record has the frequent words, the system will mark that record and search the document's name (pagoda) and message the user that pagodas have the nearly same information. System flow diagram (fig1) is shown as follow.

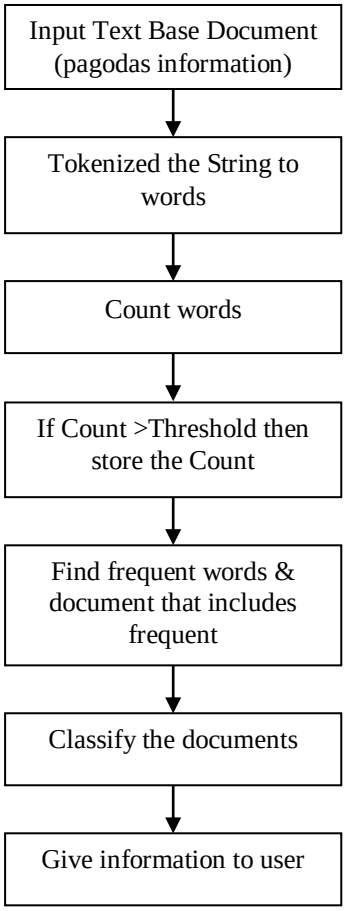

Fig 1: General Process of the system

6. Experiments and Results

A few experiments have been done for information retrieval from text based document for famous Bagan pagodas. A PC with Pentium Dual-Core E5200@2.52 GHz processor and 250 MB memory is used for our experiments. To retrieve information from many pagodas document and to need the most frequent words to classify the pagodas document, JDeveloper 11 is applied. In the experiments, most frequent words are got by using Apriori Algorithm. Figure 2 shows choosing the pagodas to classify. And then, Figure 3 and 4 shows return message after classified the selected pagodas. Searching the information by the king's name is shown in figure 5.

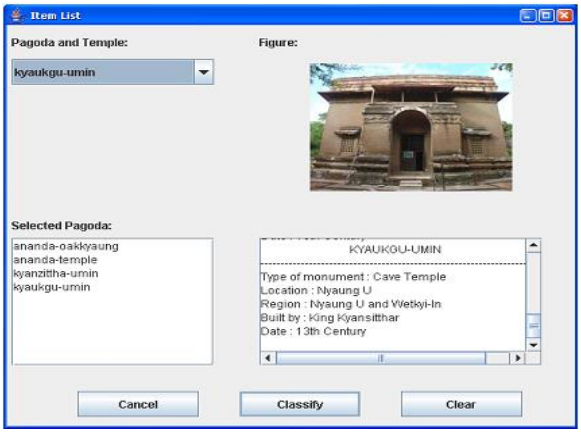

Fig 2: Choosing the pagoda to classify

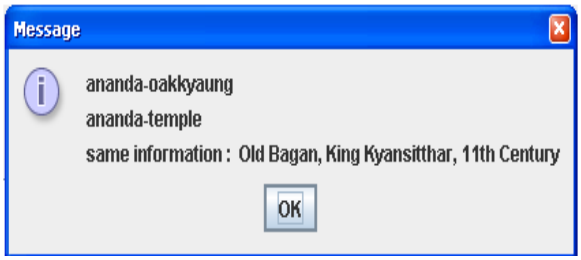

Fig 3: Returned message after classified the selected pagoda by using Apriori Algorithm

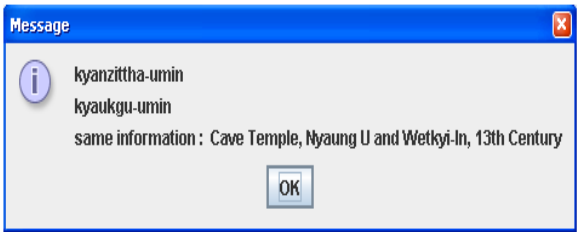

Fig 4: Returned message after classified the selected pagoda by using Apriori Algorithm

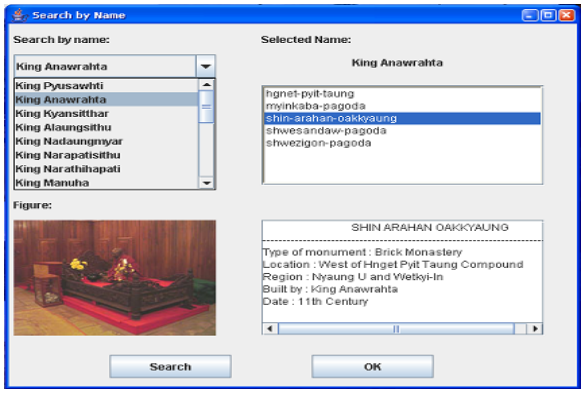

Fig 5: Searching the information by the king's name

7. Conclusion

This paper has presented a text mining technique for automatically extract association rules from collection of documents based on the keyword features. The system can be applied on all or specific parts of documents. In addition, it is designed to automatically index documents by labeling each document by a set of keywords.

In addition, association rule mining is only evaluated based on the criterion of how relevant the selected features are to the document dataset. To have a more accurate evaluation of the approach, further research should be performed on its effectiveness to the later data mining process such as categorization. The evaluation of the categorization process could then be used to evaluate the effectiveness of different feature selection approaches for text mining.

This system supports process of mining knowledge and concepts extraction from the documents. It can also provide no more time to retrieve the most important data and no more time to classify the document. The discovery of interesting association relationship among huge amount of text based document can help in decision making. Moreover, it can provide documents clustering and classification that are very useful and efficient for Information Retrieval system.

8. References

[1] Agrawal R, Srikant R.
"Fast Algorithms for Mining Association Rules", ISBN 1-55860-153-8.

[2] Chen, Xin; Wu, Yi-Fang
"Personalized Knowledge Discovery: Mining Novel Association Rules from Text"
http://www.siam.org/proceedings/datamining/2006/dm06_067chenx.pdf

[3] Ismail, Nabil; Mahgoub, Hany; Rösner, Dietmar and Torkey, Fawzy
"A Text Mining Technique Using Association Rule Extraction"
www.waset.org

[4] Kusiak, Andrew
"Association Rule the Apriori Algorithm"
http://www.icaen.uiowa.edu

[5] Nahm, Yong, Un
"Text Mining With Information Extraction"
August, 2004

[6] Written, H, Ian
"Text Mining"
http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf

[7] Written, H, Ian; Don, J, Katherine; Dewsnip, Michael
"Text Mining in a digital Library"
http://www.dcs.shef.ac.uk/~valyt/download/greenstone-gate.pdf.

[8] Wong, Chung, Pak; Whitney, Paul; Thomas, Jim
"Visualizing Association Rules for Text Mining"
http://infoviz.pnl.gov/pdf/InfoVis1999Association.pdf