Using Association Rule Techniques to Improve Document Retrieval

Timon C. Du, Jacqueline Wong, and Honglei Li
Department of Decision Sciences and Managerial Economics
The Chinese University of Hong Kong

Abstract

Document management is one of the fastest growing areas of information management. Most conventional approaches assign keywords to documents so that documents can also be retrieved through keywords. The management of data and documents are different. For example, documents have to be appeared in the full contents unlike the selected attributes are retrieved in the data queries. Also, the documents may not relevant even the values of specified attributes are identical. Moreover, documents may be related even they do not have the same keyword value. This study uses an associated knowledge rule algorithm, revised from Apriori algorithm, and classification scheme to discover the relationship between keywords of documents. The proposed algorithm can retrieve related documents without matching specified keywords in users' queries.

1. Introduction

Document management is one of the fastest growing areas of knowledge management. The end-user today is involved in saving, searching, scanning, routing, and revising documents, as well as choosing a system. According to O'Mears, a dynamic and intelligent document system is needed in order to resolve the challenging business environment and support decision making of good quality [1].

Document management merges with knowledge management at the level of the enterprise. Generally speaking, organizational knowledge system delivers the right knowledge to the right person at the right time in the right format to enable the right action [2]. Therefore, the role of conventional database administrators has been changed into knowledge administrators. The knowledge administrators manage not only the data but also the knowledge. The management of data and documents are not exactly the same, although both of them can be managed by relational database. To manage documents, several keywords are chosen for each document beforehand. The keywords are maintained by relational database. When users query document by specifying keywords. The system maps the specified keywords with documents, and the documents with matching keywords will be retrieved. Like database, knowledge management has its own terminology, ontologies, which define the shared vocabulary used in the knowledge management system to facilitate communication, search, storage, and representation [6]. However, the management of data and documents are different. For example, the document has to be appeared in the full contents unlike the selected attributes in the data queries. Also, the documents may not relevant even the values of specified attributes are identical. Moreover, documents may be related even they do not have the same keyword value. This study will propose an algorithm, called associated knowledge rule algorithm, to discover the association of keywords. So that when documents are extracted from database, some related documents would also be retrieved according to the association rule even the related documents do not contain the specified keywords. This can improve the retrieving process for transactions.

2. Data Mining And Document Classification

2.1 Data Mining

Data mining, also called knowledge discovery, is a technology of finding information from data. With the universal usage of computers more data and documents are accumulated in the database. And then, data mining techniques become a useful tool to discover the knowledge from database. There are three types of data mining problems: classification problems, association problems, and sequences problems. The classification problems group the data into clusters while association problems discover the relationship between data. The sequence problems focus on the apparent sequence of data [3]. For solving these problems, there are several well-known data mining techniques, such as classification rules, discriminant rules, clustering analysis, characteristic rules, association rules, sequences search, and mining path traversal [4]. Other than that, machine-learning approaches are also being adopted. The examples include neural network, genetic algorithm, simulated annealing [5].

This research will use the association rule approach to demonstrate the characteristics of managing documents. Several famous algorithms of the association rule approach are: Apriori [3], DHP (Dynamic Hash Pruning) [4], AIS (Agrawal, Imielinski and Swami) [7], Parallel Algorithms [8], DMA (Distributed Mining of Association Rules) [9], SETM (Set-Oriented Mining) [10], and PARTITION [11]. Since the Apriori is the most illustrious, it will be used in this study. Basically, the Apriori algorithm generates higher-level of strong related entities (documents in this study) from lower-level strong related entities. It first scans database and produces single

entities from the very primal data (transaction database in this study) and select candidate entities according to the requirement of predefined support in the first round. The support is a constant to indicate the frequencies of the occurring patterns in the data set [4]. Only the entities have larger frequencies than the support value can be considered as strong related entities. And then it will produce a table of 2-entity sets in the second round based on the strong related entities of the first round. The process is repeated until different levels strong related entities are completed obtained. The association rules are acquired from different levels of strong related entities. There are several researches dedicated to improve the efficiency of the algorithm. For example, DHP [12] use a hashed table to filter the candidate entities in each round and reduce the time of counting the support of each round.

2.2 Classification Scheme

Due to advances in storage technology, large-scale full-text retrieval systems are available at a reasonable price. The database therefore maintains not only pure data but also full-text documents. The system accepts user search queries toward both data and documents. One way to organize a document database for a full-text retrieval system is to classify a document under one or more *classes* according to the topical domains that the document discusses. This is commonly referred to as *classification*. Traditionally, classification is done by human classifiers and therefore is slow to update and operate. Also, the classification results are highly dependent on the subjective opinions and experience of a human classifier. Fortunately, automatic classification attempts to replace human classifiers by having computers to analyze the content of a document and to assign the document to the appropriate class or classes. Automatic classification has two major components: classification scheme and classification algorithm. The classification scheme defines the available classes under which a document can be classified and their inter-relationships can be specified. On the other hand the classification algorithm defines the rules and procedures for assigning a document to one or more classes defined in the classification scheme.

Most of time the classification is based on keyword subjects of documents or system vocabulary. The keywords are normally the key terms of documents while the system vocabulary consists of subject thesaurus. Most computer systems adopt both the controlled vocabulary and keywords as the subject terms to represent the documents. Moreover, information such as author's name, date of publication, and language can also be used for searching. The indexed documents are kept in the document storage while the document representations—that is, surrogates such as keywords, vocabulary, and so forth—are used for matching. Both the documents and surrogates are stored in the database for future searches. Using this approach, the retrieval process will be very efficient.

3.Using Association Rules to Discover Knowledge

This study argues that discovering knowledge from documents is different from data in the following ways:

1. The document has to be appeared in the full contents unlike the selected attributes in the data queries.
2. The documents may not relevant even the values of specified attributes are identical. For example, two documents having value of *virus* in a keyword attribute do not mean that they are in the same category. In fact, one document may belong to computer science and the other can be an article published by Department of Health.
3. Both SQL keywords search and full-text browsing can be used to search documents.
4. The Boolean operation can be used to filter the queried documents in exact match, but is not appropriate for mapping to related terms.
5. The association rule can be used to discover the documents even without specified keywords. Referring the classification scheme as described in session 2 can do this.
6. The classification scheme can be updated by association rules.

Section 3.1 describes the framework of this study while using Apriori algorithm is presented in session 3.2. An algorithm for applying association rule is discussed in session 3.3.

3.1 The Framework of Document Knowledge Discovery

Most of current system maintains documents in the relational database. The documents are allowed to assign numeric keywords and the keywords are used to retrieve documents. The keywords of documents are constructed into relations of relational database in the keyword database. The users type-in segmented words as anchored to query documents. During query processing, Boolean operation can be further used to limit the query results queried by the segmented words. However, the conventional approach cannot retrieve a document without specified keywords. This is why association rule technique can improve the knowledge discovery of documents. The association rule identifies the occurrences of entities and discovers the relationship between entities. The Apriori algorithm of association rule techniques analyzes the transactions and determines the strong related entities. However, as will be shown in the example of session 4, some irrelevant documents may be erroneously identified. For example, the virus can infect both human and computers. But, it will be considered as two different subjects if we are discussing about virus infection. In this case, the classification scheme is used to recognize the

subjects of keywords. The keywords are classified into different subjects classes in the classification scheme. The entities belonged to different subjects should not be retrieved as the related document. That is, the classification scheme can distinguish the difference between computer virus and virus disease. Then, the association rule is memorized as associated knowledge rules. Therefore, combining both associated knowledge rules and classification scheme, the documents having matched keywords and related words will be returned to users.

3.2 Mining Association rules and Apriori algorithm

Given a knowledge database, the association rules search for the relations between ontologies (keywords of documents) such that the presence of some ontologies will imply the presence of other ontologies when the same knowledge (document) are retrieved. Let $K = \{k_{11}, k_{12}, ..., i_{1n}, k_{21}, k_{22}, ..., i_{nm}\}$ as the set of ontologies. Let D_i of D be a set of retrieved objects where $K_{ij} \subset D_i$. Each transaction t_i in T is represented as a set of ontologies such that $T \subseteq D$. Let A be a set of objects and a transaction t_i is said to contain A if and only if $A \subseteq T$. Also, let X be a set of ontologies and a transaction t_i is said to contain X if and only if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset T$, $Y \subset T$, and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ holds in the transaction set T with confidence c if c% of retrieves in T that contain X also contain Y. The rule $X \Rightarrow Y$ has support s in the transaction set T if s% of transactions in T contain $X \cup Y$. Note that the confidence denotes the strength of implication and the support indicate the occurring patterns in the rule [4]. Apriori algorithm scans the transaction database and counts the occurrences each entity and generates a 1-level document relationship. Then the occurrences are filtered by the predefined support. Then it generates 2-level documents according to the Cartesian product $L_k * L_k = \{I_i \cup I_j \mid I_i, I_j \in L_k, |I_i \cap I_j| = k-1\}$, where *k* is the entities level, L_k is the strong entities of the *k* level and I_i, I_j are the entities of L_k. Then it filters 2-level entities according to the predefined support.

3.3 Associated Knowledge Rules

The associated knowledge rules are obtained from applying the association rule technique toward both the document database and transaction database. The Apriori algorithm of the association rule techniques is adopted in this study. This algorithm evaluates the occurrences of entities (documents) and only the occurrences over the predetermined support value will be adopted. The support value specifies the minimum strongly level of entity relationship. The Apriori algorithm begins with finding the occurrence of one entity and then reaches occurrence of higher levels entity. To simplify the problem illustration, the algorithm of associated knowledge rules only adopts 2-entity generation. The higher degree can be implemented in the similar manner.

5. Conclusion

This study revised the Apriori algorithm of association rules into an associated knowledge rule algorithm. The algorithm can discover the relationship between keywords of documents. By using both the associated knowledge rules and classification scheme, the related documents without specified keywords can also be retrieved.

References

[1] O'MEARA, D., Oct. 2000, Buried in documents? *Engineering Management Journal*, 10(5), 241 -243.

[2] DIENG, R., Knowledge management and the internet, May-June 2000, *IEEE Intelligent Systems*, 15(3), 2000, 14 -17.

[3] AGRAWAL, RAKESH, IMIELINSKI, TOMASZ, and SWAMI, ARUN, DEC 1993, Database Mining: A Performance Perspective, *IEEE Transaction on Knowledge and Data Engineering*, 5(6), 914-925.

[4] CHEN, MING-SYAN, HAN, JIAWEI, and YU, PHILIP S., DEC. 1996, Data Mining：Overview from a Database Perspective, *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 866-883.

[5] BERRY, MICHAEL J. A. and LINOFF, GORDON S., 1997, *Data Mining Techniques, For Marketing, Sales, and Customer Support*, (New York: John Wiley & Sons, Inc.)

[6] Daniel E. O'Leary, 1998, Enterprise Knowledge Management, Computer, Volume: 31 Issue: 3, March 1998, P54-61

[7] AGRAWAL, RAKESH, IMIELINSKI, TOMASZ, and SWAMI, ARUN, MAY 1993, Mining Association Rules between Sets of Items in Large Databases, *Proc. of the 1993 International Conference on Management of Data (SIGMOD-93)*, 207-216.

[8] AGRAWAL, RAKESH and SHAFER, JOHN C., DEC 1996, Parallel Mining of Association Rules, *IEEE Transaction on Knowledge and Data Engineering*, 8(6), 962-969.

[9] CHEUNG, DAVID W., NG, VINCENT T., FU, ADA W., and FU, YONGJIAN, DEC 1996, Efficient Mining of Association Rules in Distributed Databases, *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 911-922.

[10] HOUTSMA, MAURICE and SWAMI, ARUN, JUL 1995, Set-oriented Data Mining in Relational Databases, *Data and Knowledge Engineering*, 17, 245-262.

[11] AGRAWAL, RAKESH, MANNILA, SRIKANT, RAMAKRISHNAN, TOIVONEN, HANNU, and VERKAMO, A. INKERI, 1996, Fast Discovery of Association Rules, *Advances in Knowledge Discovery and Data Mining*, *AAAI press*, 307-328.

[12] J-S Park, M. –S Chen, and P. S. Yu, 1995, An Effective Hash Based Algorithm for Mining Association Rules, Proc. ACM SIGMOD, 175-186.