INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT
IJARnD
(Volume 5, Issue 5)
Available online at: www.ijarnd.com

Knowledge discovery from documents by using Text Mining Technique

Ni Ni Khaing
University of Computer Studies, Meiktila, Myanmar

ABSTRACT

Text mining is an important field because of the necessity of obtaining knowledge from the enormous number of text documents. This system describes text mining technique for automatically extracting association rules amongst keywords from collections of textual documents. This system integrates XML technology with Information Retrieval scheme (TF-IDF) for keyword selection that automatically selects the most discriminative keywords for use in association rules generation and uses data mining technique for association rules discovery. This system consists of text preprocessing phase for filtration and indexing of the document, knowledge distillation phase for generating association rules by using GARW algorithm and visualisation phase for displaying results. This system extracts association rules that contain important features and describe the informative news (knowledge) that are included in the XML documents collection by using generated association rules. This system is implemented by using C# programming language.

Keywords— *Text Mining, Association Rule Mining*

1. INTRODUCTION

Knowledge Discovery in Texts, Text Mining (TM), is a new research area that tries to solve the problem of information overload by using techniques from data mining, machine learning, information retrieval and knowledge management.

Association rules highlight correlations between keywords in the texts. A word is selected as a keyword if it does not appear in a predefined stop-words list. Moreover, association rules are easy to understand and to interpret for an analyst. This paper focus on the extraction of association rules amongst keywords labeling the documents.

The problem of text mining is that unlike tabular records in databases, documents are not structured. Therefore, computers could easily recognize them. The lack of explicit structure raises the difficulty of uncovering the implicit knowledge inside the documents. It is hard to extract and represent abstract concepts from a natural text. The emerging standard XML and its supporting techniques help to structure and automate indexing of document. This makes part of the semantics of a document explicit and thus machine processable.

The main contribution is that this paper uses an Information Retrieval scheme (TF-IDF) to extract keywords within XML documents and then data mining technique is also used to extract knowledge within these documents.

2. RELATED WORKS

I.H. Witten [1] applied existing data mining techniques to discover episode rules from text. Episode rule mining is used for language analysis because it preserves the sequential structure of terms in a text document. However, extraction of association rules that get the relations of the existing of the keywords in text ignoring the order in which these keywords occur.

M. Rajman and R. Besancon [5] presented two examples of text mining tasks and prototypical document extraction, along with several related NLP techniques. In association extraction task, they had extracted association rules from indexed documents collection. Finding information in a collection of indexed documents by automatically retrieving relevant associations between keywords was also presented in [5].

3. BACKGROUND THEORY

3.1. Text Mining

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 is the core process of knowledge discovery in text documents. It is the data analysis of text resources to that new, previously unknown knowledge is discovered [6]. Text mining framework is shown in Figure 1.

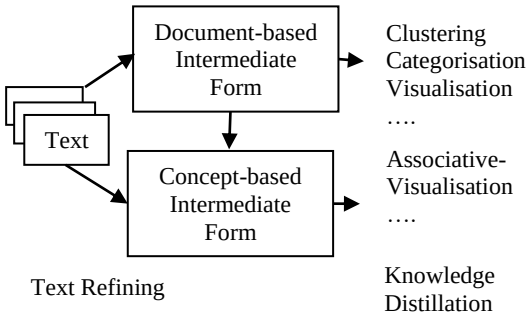

Fig. 1: Text Mining Framework

3.2. Text Preprocessing

Some preprocessing tasks are usually performed before the documents in a collection are used for retrieval. For traditional text documents, these documents are filtered to eliminate the unimportant words by using a list of stop words. The filtered documents are then indexed by using the weighting (TF-IDF) scheme. If the textual data is indexed, the indexing structures can be used as a basis for actual knowledge discovery process [2].

3.2.1 Filtration: In this process, documents are filtered by removing the unimportant stopwords from document content. Stopwords removal process improves information retrieval and searching by ignoring words that usually appear in every document. Thus, it is not helpful in distinguishing documents from each other. Stop words are not useful for text mining [3]. Additionally, the removal of stopwords reduces the index size (number of distinct words in the index) and therefore save space and time [4].

3.2.2 Indexing: Techniques for automated production of indexes associated with documents can be borrowed from the Information Retrieval field. Each document is described by a set of representative keywords called index terms. It is obvious that different index terms have varying relevance when used to describe document contents in a particular document collection. This effect is captured through the assignment of numerical weights to each index term of a document.

The techniques usually rely on frequency-based weighting schemes. The weighting scheme TF-IDF is used to assign higher weights to distinguish terms in a document, and it is the most widely used weighting scheme, which is defined as:

$$w(i, j) = tfidf(Ndi, tj) = Ndi, tj * \log 2 |C|/Ntj$$

where, in the term frequency (tf) factor, Ndi, tj denotes the number of the term tj that occur in the document di. In the inverse document frequency (idf) factor, Ntj denotes the number of documents in collection C in which tj occurs at least once and |C| denotes the number of the documents in collection C.

Once a weighting scheme has been selected, indexing can be performed for each document by simply selecting the keywords that satisfy the given weight constraints. Weight constraints is used to identify and filter the keywords that may not be of interest in the context of the whole document collection either because they do not occur frequently enough, or they occur in a constant distribution among the different documents [4].

3.3. Knowledge Distillation

Knowledge distillation phase presents a way for finding information from a collection of indexed documents by automatically extracting association rules from them. In this system, knowledge is distilled using the GARW [2].

3.3.1 Association Rule Mining: Association rule mining finds interesting association or correlation relationships among a large set of keywords. There are two important basic measures for association rules, support (s) and confidence (c). The rule $W_i \Rightarrow W_j$ has support s in the collection of documents D if s% of documents in D conation $W_i \cup W_j$. The support is calculated by the following formula:

$$Support(W_i W_j) = \frac{Support Count\ of\ W_i W_j}{Total\ number\ of\ documnets\ D} \quad (1)$$

The rule Wi ⇒ Wj holds in the collection of documents D with confidence c if among those documents that contain Wi, c % of them contain Wj also. The confidence is calculated by the following formula:

$$Confidence(W_i | W_j) = \frac{Support(W_i W_j)}{Support(W_i)} \quad (2)$$

An association rule-mining problem is broken into two steps:

- Generate all the keyword combinations (keywordsets) whose support is greater than the user specified minimum support (called minsup). These sets are called the frequent keywordsets and
- Use the identified frequent keywordsets to generate the rules that satisfy a user specified minimum confidence (called minconf) [4].

3.3.2 Generating Association Rules based on Weighting scheme (GARW) Algorithm: GARW algorithm scans only the generated XML file during the generation of the large frequent keyword sets. The GARW algorithm is as follows:

1. Let *N* denote the number of top keywords that satisfy the threshold weight value.
2. Store the top *N* keywords in index XML file along with their frequencies in all documents, their weight values TF-IDF and documents ID. Four XML tags for all keywords (<doc-id>, <keyword>, <keyword-frequency>, <TF-IDF>) index the file.
3. Scan the indexed XML file and find all keywords that satisfy the threshold minimum support. These keywords are called large frequent 1-keywordset L_1.
4. In $k \geq 2$, the candidate keywords C_k of size *k* are generated from large frequent (*k*-1)-keywordsets, L_{k-1} that is generated in the last step.
5. Scan the index file, and compute the frequency of candidate keyword sets C_k that generated in step 4.
6. Compare the frequencies of candidate keyword sets with minimum support.
7. Large frequent *k*-keyword sets L_k, which satisfies the minimum support, is found from step 6.
8. For each frequent keyword set, find all the association rules that satisfy the threshold minimum confidence [2].

3.4. Visualisation Process

The extracted association rules are reviewed in textual format or tables, or in graphical format. This process is designed to visualise the extracted association rules in textual format or tables [4].

4. PROPOSED SYSTEM ARCHITECTURE

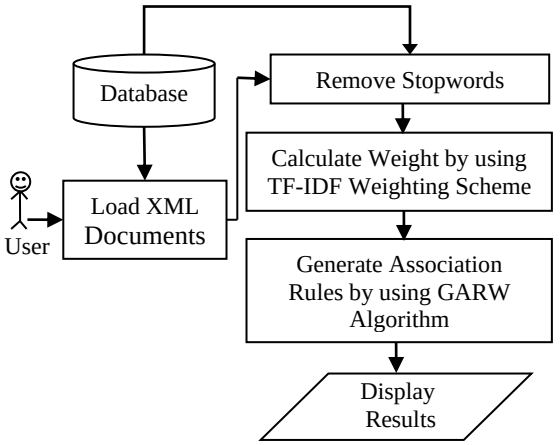

Fig. 2: Proposed System Architecture

This system automatically discovers association rules from XML documents. Firstly, XML documents are loaded from the database. And then, these documents are filtered to remove the unimportant stopwords from documents content by using the stopwords list database. The filtered documents are then indexed by using TF-IDF weighting scheme after the filtration process. TF-IDF is applied to select significant noun phrases from each target document. And then, this system finds information from a collection of indexed documents by automatically extracting association rules using GARW algorithm. After generating association rules, this system displays these rules as a result to the user. Proposed system architecture is shown in Figure 2.

4.1. System Flow Diagram

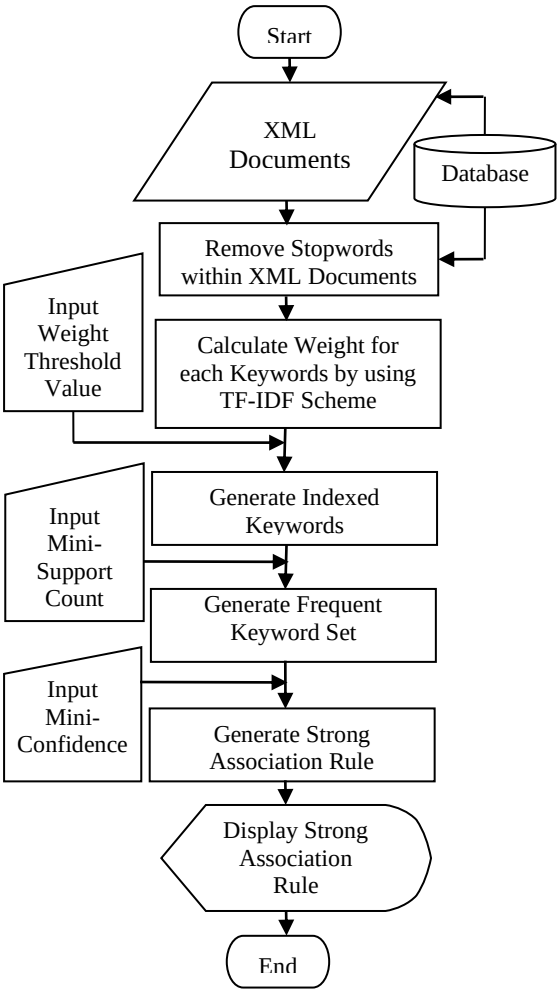

Fig. 3: System Flow Diagram

In system flow diagram shown in Figure 3, firstly, XML documents are loaded into the system. And then, stopwords within these XML documents are filtered by using the stopwords list in the database. The stopwords list consists of unimportant words such as a, an, the, etc. The words within filtered XML documents are keywords of these XML documents. And then, this system calculates the weight values for each keyword in all documents. And then, the user must define any desired weight value. The system generates the indexed keywords according to the user defined weight value.

After generating keywords, the user must define the minimum support count. In this system, minimum support count is used to know the count of keywords. The more the count of keywords within XML documents, the more the importance of these keywords. Then, the system is compared with the minimum support count. Keywords less than minimum support count is removed and other going on processing. After finishing these processing, the system generates frequent keyword set in all documents. Finally, the rules having equal to or greater confidence than user specified one, are considered to be strong association rules.

4.2. Knowledge Discovery in XML Document

This system is implemented to extract association rules depending on the analysis of relations between the keywords in the documents collection about "Neurological Diseases". Finding rules in text documents are useful in a number of contexts such as investigations, and in general understanding affect of events in the real world.

In this paper, we use three XML documents about "Neurological Diseases" to implement knowledge discovery in documents by using GARW algorithm. These documents are:

XML Document 1 (D1)

<?xml version = "1.0" encoding = "UTF-8"?>
<name = "disease">
<description>
Headache is a frequently encountered neurological symptom but is seldom associated with significant neurological disease.
</description>
</name>

XML Document 2 (D2)

<?xml version = "1.0" encoding = "UTF-8"?>
<name = "disease">
<description>
Patients usually fear serious brain disease.
</description>
</name>

XML Document 3 (D3)

<?xml version = "1.0" encoding = "UTF-8"?>
<name = "disease">
<description>
Patients with Headache who are normal on neurological disease examination are unlikely to have a serious disorder, however distressing their symptom.
</description>
</name>

In text preprocessing, these documents are filtered by removing the stopwords from documents content. This system removes stopwords within documents by using the stopwords list in the database. The words that are useless in text mining are called stopwords such as articles (a, an, the), prepositions (in, at, on, about) and so on. After removing stopwords, weight values are calculated for each keyword in each document. The words within filtered documents are keywords of these documents. The filtered documents are the documents, which have been removed stopwords. This system can also extract keywords from complex and not similar format documents by removing stopwords. Table 1 shows weight value of some words within documents.

Table 1. Weight Value of Some Words

Term	Nd_i, t_j			$\log_2 \|C\|/Nt_j$	w (i, j)		
	D1	D2	D3		D1	D2	D3
Headache	1	-	1	0.58	0.6	-	0.6
frequently	1	-	-	1.58	1.6	-	-
encountered	1	-	-	1.58	1.6	-	-

neurological	2	-	1	0.58	1.2	-	0.6
symptom	1	-	1	0.58	0.6	-	0.6
seldom	1	-	-	1.58	1.6	-	-
associated	1	-	-	1.58	1.6	-	-

In the association rule mining phase, this system presents a way for finding information (knowledge) from a collection of indexed documents by automatically extracting association rules from them.

In generation of indexed keywords, the user defines the threshold weight value as 0.4. So, the document database consists of document ID and keywords that are satisfied the user defined weight value. Keyword table is shown in Table 2.

Table 2. Keyword Table

XML Document ID	Keywords
D1	Headache, frequently, encountered, neurological, symptom, seldom,
D2	associated, significant
D3	Patients, fear, serious, brain
	Patients, Headache, normal, neurological, disease, examination, serious, disorder, distressing, symptom

Table 3. Frequent 1-Keywordsets

KeywordSet	Support Count
{Headache}	2
{neurological}	2
{symptom}	2
{Patients}	2
{serious}	2

User defines the minimum support count is 2. After counting their supports, the frequent keyword sets are generated by the GARW algorithm. The keyword set that appears in fewer than minimum support count 2 is discarded. Frequent keyword sets are shown in Table 3, 4 and 5.

Table 4. Frequent 2-Keywordsets

KeywordSet	Support Count
{Headache, neurological}	2
{Headache, symptom}	2
{neurological, symptom}	2
{Patients, serious}	2

Table 5. Frequent 3-Keywordsets

KeywordSet	Support Count
{Headache, symptom, neurological}	2

In the visualisation phase, association rules come out as output analytical result after generating frequent keyword sets. Then, the user must define the minimum confidence as the threshold value to generate the strong association rule. The rules, having equal to or greater confidence than user specified one, are considered to be strong. Even with the minimum confidence 100%, the implemented system can extract the strong rules for the given XML documents. The experimental results are shown as textual format.

The extracted association rules are as follows:

Rule 1: Headache →symptom, neurological
confidence= (2/2) * 100 = 100%
Rule 2: symptom →Headache, neurological
confidence= (2/2) * 100 = 100%
Rule 3: neurological→Headache, symptom
confidence= (2/2) * 100 = 100%
Rule 4: symptom, neurological → Headache
confidence= (2/2) * 100 = 100%
Rule 5: Headache, neurological → symptom
confidence= (2/2) * 100 = 100%
Rule 6: Headache, symptom → neurological
confidence= (2/2) * 100 = 100%

This system extracts more relationships between keyword from each document. According to the rule 1, this system extracts the relationship between Headache, symptom and neurological. So, the user can know the knowledge about Headache keyword associated with symptom and neurological keywords. Moreover, the user can also know Headache, symptom and neurological that are main information within three documents. In the real world, the association rules that are extracted from the documents can be applied to easily obtain and understand the essence within these documents for the user.

5. EXPERIMENTAL RESULTS

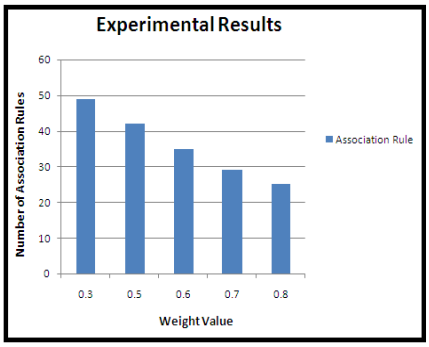

Fig. 4: Experimental Results

This system is tested with 200 XML documents about neurological diseases. These documents are retrieved from the "Davidson's Principles and Practice of Medicine, 20th Edition" book in the medical field. This system is proposed for the analysis of XML documents using association rule mining by analyzing the keyword sets. According to support and confidence, this system generates association rules that are the results of analysis report by using GARW algorithm. The results can be changed according to the user threshold value of weight, support count and confidence value. Figure 4 shows experimental results of this system. These results are obtained by using 50 XML documents, support count value is 2 and confidence value is 100%. Figure 4 shows experimental results.

6. CONCLUSION

The system is proposed to extract association rule from XML documents. This system finds the relations between keywords and presents them in association rules form. And then, this system gives useful information (knowledge) from these association rules to the user about domain. The extracted useful information from the XML document is based on the abstractions that describe the relationships between the keywords in texts. To investigate the use of GARW algorithm, this system applied this algorithm on selected sample of XML documents that are related to the Neurological Disease.

7. REFERENCES

[1] I.H. Written, "Adaptive Text Mining: Inferring Structure from Sequences," *Journal of Discrete Algorithms*, vol. 2, no. 2, pp. 137-159, University of Waikato, New Zeland, 2004.

[2] I.T. Fatudimu, A.G. Mausa, A.B. Sofoluwe, C. K. Ayo, "Knowledge Discovery in Online Repositories: A Text Mining Approach", *European Journal of Scientific Research*, vol. 22, pp. 241-250, 2008.

[3] L.Bing, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, .2nd edition, Springer, July 2011.

[4] M.Hany, R.Dietmar, I.Nabil and T. Fawxy, "A Text Mining Technique Using Association Rules Extraction", *Int. Journal of Information and Mathematical Sciences*, 2008.

[5] M.Rajman and R.Besancon, "Text Mining: natural language techniques and text mining applications", in *Proc. Of the 7th IFIP Working Conference on Database Semantics*, pp. 7-10, Switzerland, October 1997.

[6] S. Martin, B. Peter and P. Jan, "Grid-based Support for Different Text Mining Tasks", *Acta Polytechnica Hungarica*, vol. 6, no. 4, 2009.

