Proc. of 2nd International Conference on Advances in Computer, Control & Telecommunication Technologies (ACT 2010), pp.122-124

Intelligent Searching using Association Analysis for Law Documents of Indonesian Government

Iis Siti Darawaty, Siti Syarah
Department of Information Technology, Faculty of Science & Technology
Universitas Al Azhar Indonesia Komplek Masjid Agung Al Azhar
Jakarta Indonesia 12110
Email: i2zdw47@gmail.com

Anto Satriyo Nugroho, Fara Ayuningtyas, Yuki Istianto, Bowo Prasetyo, Mohammad Teduh Uliniansyah, Made Gunawan, Desiani, Asril Jarin, Dwi Handoko
Center for Information & Communication Technology, Agency for Assessment & Application of Technology
BPPT 2nd building 4F, Jalan M.H.Thamrin 8
Jakarta, Indonesia 10340
Email: asnugroho@ieee.org

***Abstract*—Information Retrieval (IR) technology helps people to search relevant and necessary documents from massive digital database. In the context of e-law enforcement, it can be used to find verses related to a certain topic and presenting the relationship between one verse to the others. In this study we propose an intelligent searching system for Indonesian law documents which is enhanced by association analysis to discover the association of law related keywords, thus providing guidance for the user to find related verses.**

***Keywords*—association analysis, law documents, intelligent searching, information retrieval**

I. INTRODUCTION

Information Retrieval technology helps people to find relevant and necessary documents from massive digital database. Once a document has been digitalized, it becomes easy for users to find relevant information from the database. Google.com, yahoo.com are examples of such service which have a great contribution to provide online information quickly and accurately to the internet users.

Various datamining algorithms are incorporated, since they are capable to discover knowledge from the database which can be used to increase the searching performance. One of such techniques is Association Analysis. Examples of its application are found in the field of webmining[1], document analysis[2], and bioinformatics[3].

In this study, we developed "Telusur Hukum", an information retrieval system to help people finding Indonesian law documents and verses related to a certain topics and show the association between one verses to the others. The association among the verses were discovered using *Apriori* algorithm in Association Analysis[4].

The structure of this paper is organized as follows: Sec.2 described the "Telusur Hukum" system, Sec.3 described the application of Association Analysis to discover the associations between keywords in law documents, Sec.4 reported the experimental results which will be summarized and concluded in Sec.5.

II. TELUSUR HUKUM

"Telusur Hukum", an information retrieval system to help people finding Indonesian government law documents. The complete system would have a database contain the law documents, Indonesian language WordNet, a set of applications for searching keywords (law cases), viewing the clusers of the documents, and visualizing the searching results in a concise and easy to understand manner. Currently, the database contains more than 2100 verses from 28 law and ordinance documents. This system can be accessed from http://www.inn.bppt.go.id/TelusurHukum/index.php

This paper reported the efforts to enhance the searching system by providing related keywords which were discovered using association analysis.

III. MINING THE LAW DOCUMENTS USING ASSOCIATION ANALYSIS

Intelligent Searching system aims to help user not only finding documents that match to a specific query, but also providing candidates or documents which are considered closely related to the query of interest. To develop an intelligent searching system,various datamining techniques are implemented to reveal the knowledge inside the database. Various algorithms are evaluated to enhance the searching capability of "Telusur Hukum", including Association Analysis, Clustering on topological visualization based on Kohonen's Self Organizing Maps. This focus of this paper reported the evaluation of Association Analysis implemented in Telusur Hukum.

Figure 1 shows the overview of the proposed Association Analysis based intelligent searching system. Due to the unstructure characteristics of law documents, it is necessary to proceed the data by converting them into vectoral representation, thus ready to be processed by Association Analysis. It consists of tokenization, stop word removal, text to vector conversion, association analysis using Apriori algorithm, then association rules interpretation. In the tokenization phase, the sentences will be broken into words (tokens), using tab and

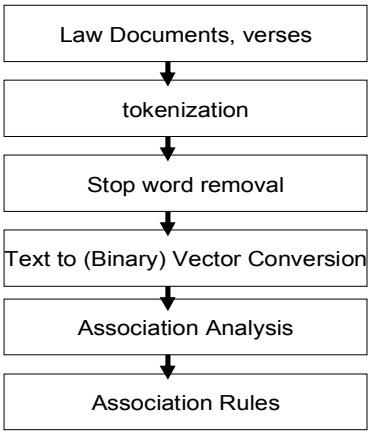

Figure 1. Overview of the system

characters spaces as delimiters. A list of Indonesian stop words are used to remove those words from the documents to increase the search performance. One critical step in this study is choosing the appropriate vector representation of the verses. A verse or document can be represented using in various levels: sub-word level, word-level, multi-word level, semantic level and pragmatic level. We chose word-level representation due to its simplicity. The attribute of the vector could be binary, raw term frequency, verses, document frequency, or their combination. The current system adopted the binary representation, by assigning "1" if a word found in a verses, and "0" otherwise [5]. After converting the verses of law documents into vector representation, the associations of the keywords are extracted using Association Rules.

Association rules are formally expressed as $X \rightarrow Y$, where *X* and *Y* are disjoint itemsets. In the experiments, the itemsets are the keywords found in the law documents. The strength of the rule is measured in terms of support *s* and confidence *c*. They are formally defined as follows:

$$\text{Support, } s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}; \quad (1)$$

$$\text{Confidence, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}. \quad (2)$$

Support *s* is a metric that shows how often a rule is applicable to a given dataset, while confidence *c* shows how frequently keyword in *Y* appear in the verses contain keyword *X*.

Thus, the aim of this step is to extract rules from the database, i.e. *"if keyword-1 found in a verse, keyword-2 will found in the same verse with support s and confidence c"*. These rules will help user to find verses related to a topic by providing several combinations keywords (*keyword-1* and *keyword-2*) of which their associations have been discovered from the database at certain values of support *s* and confidence *c*.

Association Analysis basically consists of two major subtasks:

1. Frequent Itemset Generation, to find all itemsets (keywords) that satisfy a minimum threshold
2. Rule Generation, to extract all high confidence rules from the frequent itemsets.

Frequent itemset generation is implemented using *Apriori* Algorithm [4][6]. A*priori* used support-based pruning to systematically control the exponential growth of the candidate itemsets. Let C_k denotes the set of the candidate *k*-itemsets and F_k denotes the set of frequent *k*-itemsets:

1. The algorithm determines the support of each item (in this case : keyword), thus the set of all frequent 1-itemsets are obtained (F_1)
2. Generate new candidate *k*-itemsets using the frequent (*k*-1) itemsets found in the previous iteration
3. Count the support of the candidates
4. Eliminates all candidate itemsets whose support count are less than *minsup* (minimum support threshold)
5. The algorithm terminates when there are no new frequent itemset generated.

IV. EXPERIMENTAL RESULTS

Experiments were conducted using law documents which were obtained from Indonesian Ministry of Law and Human Rights [7]. All the documents are written in Indonesian. We used 300 verses of Indonesian government (Specifically, they consists of "Undang-undang", "Peraturan Pemerintah", "Keputusan Presiden", "Peraturan Presiden", "Undang-undang Darurat Republik Indonesia", "Peraturan Pemerintah Pengganti Undang-Undang" and "Penetapan Presiden"). The data were preprocessed by tokenization, stop word removal and vector conversion as depicted in Fig.2. Stopword list consists of 330 words, such as "jika", "dan", "yang", "namun". The experiments were conducted using computer and software with configuration as follows: Pentium IV, 2GB RAM, Windows Vista, WEKA version 3.4.13 [8].

To convert each verse of the documents into vector representation, we monitored the presence of a word in a verse instead of measuring it frequency, thus yielded binary representation. A value of 1 indicated that a certain keyword existed in the verse, while 0 indicated the absence of the keyword. A list of 125 keywords were generated manually which were considered to have important meaning. After recorded the presence/absence of these words, a verse was converted into 123-dimensional binary vector representation.

The database was then processed using Association analysis. The possible rules extracted from the dataset with *d*=123 keywords (itemset) is around 4.85×10^{59} possible rules. Association analysis was applied to prune the rules, extracting only interesting ones from the database. The first step is Frequent Itemset Generation. A set of keyword combinations were extracted by Apriori Algorithm as frequent itemsets at various minimum support (*s*) *minsup*.

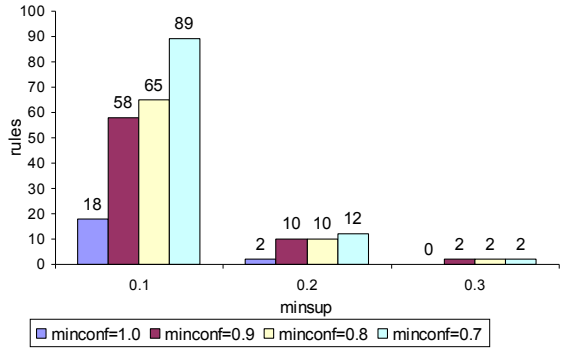

Figure 2. The number of association rules extracted in various threshold of *minsup* and *minconf*

The second step is Rule Generation, aims to extract all high confidence rules from the frequent itemsets. Figure 2 depicted the number of association rules extracted at *minsup* 0.1, 0.2, 0.3 and *minconf* 0.7, 0.8, 0.9 and 1.0.

Based on the result, we chose *minsup*=0.1 and *minconf*=0.9, to obtain a sufficient number of rules with high level confidence. This combination of thresholds generated 58 association rules. Table 1 shows part of the rules obtained in the experiments.

The first rule of Tab.1 (No.1) shows that from 300 verses there are around 60 verses that have keyword *pendaftaran* (registration) since the support *s* is 0.2. Confidence level of this rule *c*=1.0 means that all of the 60 verses contain keyword *penduduk* (population). This rule is one of strongest found in this experiment. The higher value of support *s* indicated the higher importance of the rules, while the higher value of confidence *c* indicated the higher reliability of the rules. The second rule (No.2) shows an example of rule which associated two keywords (*instansi* and *pendaftaran*) and one keyword *penduduk*. Although the rule has high confidence level (*c*=1.0) it has lower support compared to the first rule at *s*=0.1, thus the association rule is found in only 10 verses. The rest of the examples shows the rules obtained in various support *s* and confidence *c*.

Using the result of the experiments, the Telusur Hukum will assist user by providing associated keywords as guidance to enhance the searching. For example, if a user wrote "pendaftaran" as searching query to Telusur Hukum, the system will response by providing associated keywords information e.g. "pendaftaran" is associated with "penduduk" at *s*=0.2 and *c*=0.1 and give the links to the related verses.

V. CONCLUSIONS

This study aims to implementing Association Analysis to discover association rules among verses of Indonesian Government Law Documents. The extracted rules are used to enhance the searching function of "Telusur Hukum", an intelligent searching system for Law Documents.

TABLE 1 ASSOCIATION RULES OBTAINED AT *MINSUP*=0.1 AND *MINCONF*=0.9

No.	Keyword X→Y	s	c
1	pendaftaran →penduduk	0.2	1.0
2	instansi, pendaftaran →penduduk	0.1	1.0
3	pelaksanaan, pendaftaran →penduduk	0.1	1.0
4	dokumen, data → kependudukan	0.1	1.0
5	Administrasi→kependudukan	0.2	0.9

When a searching query is inputed to the system, it will provide associated keywords information and their links which will help the users to find related verses. Experiments were conducted using 300 verses and the top 58 rules were obtained at minimum support threshold *minsup* 0.1 and minimum confidence *minconf* =0.9. The rules helped users to find verses that have the associated keywords. The future works of this study includes conducting experiments with larger scale database, integrating the system with verses clustering based on Kohonen Self-Organizing Maps, and development of Indonesian WordNet.

ACKNOWLEDGEMENTS

This publication is supported by Universitas Al Azhar Indonesia.

REFERENCES

[1] J.Pei, J.Han, Bo.Mortazavi-Asl, and H.Zhu, "Mining Access Patterns Efficiently from Web Logs," Proc. Of the 4th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, pp.396-407, Kyoto, Japan, April 2000

[2] J.D.Holt, S.M.Chung, "Efficient Mining of Association Rules in Text Databases," Proc.of the 8th International Conference on Information and Knowledge Management, pp.234-242, Kansas City, Missouri, 1999

[3] H.Xiong, X.He, C.Ding, Y.Zhang, V.Kumar and S.R.Holbrook, "Identification of Functional Modules in Protein Complexes via Hyperclique Pattern Discovery," Proc. Of the Pacific Symposium on Biocomputing, Maui, January 2005

[4] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," 20th International Conference on Very Large Data Bases, 478-499, 1994.

[5] T. Joachims, Learning to Classify Text using Support Vector Machines, Kluwer/Springer, 2002

[6] P.N.Tan, M.Steinbach, V.Kumar, Introduction to Datamining, Addison Wesley

[7] Indonesian Law Documents published by Ministry of Law and Human Rights http://www.djpp.depkumham.go.id/ (last accessed: 1 July 2010)

[8] M.Hall, E.Frank, G.Holmes, B.Pfahringer, P.Reutemann, I.H.Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, Vol.11, Issue No.1, 2009