AENSI Journals
Australian Journal of Basic and Applied Sciences
ISSN:1991-8178
Journal home page: www.ajbasweb.com

Design of an Information Retrieval System for Malay Language Fatwa Documents

[1]S. Fahmy, [1]I, Nazri, [2]I. Evi, [3]M. Rohana

[1]*Faculty of Computer, Media & Technology, TATI University College, 24000Terengganu, Malaysia*
[2]*Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM, Serdang, Malaysia*
[3]*Faculty of Computer Science & Information Technology, University of Malay, 50603 Kuala Lumpur, Malaysia*

ARTICLE INFO

Article history:
Received 20 November 2013
Received in revised form 24 January 2014
Accepted 29 January 2014
Available online 5 April 2014

Keywords:
Information Retrieval, Vector Space Model, Fatwa, Malaysia

ABSTRACT

Fatwa is a legal pronouncement in Islam and issued by a recognized religious authority. Fatwas are issued in response to questions regarding daily matters in accordance with religious law such as gender relations or the use of new technology. In Malaysia, fatwa is under the purview of the Fatwa Committee of the National Council for Islamic Religious Affairs, overseen by the Malaysian Department of Islamic Development (JAKIM). The committee is the body issuing fatwa at the national level on matters referred to it by the Conference of Rulers. Currently, there are at least two on-line systems for retrieving Malay language Fatwa documents, e-Fatwa and Fatwa Management System provided by JAKIM and the World Fatwa Management and Research Institute respectively. Both systems provide search capabilities for fatwa documents using Boolean expressions. Due to the limitations of the query builder and search engine, users face difficulties in constructing effective queries, leading to irrelevant documents being returned. This paper proposes a design for an information retrieval system for fatwa documents based on different document structures. In the design, fatwa documents undergo the process of digitizing; stop-words removal; stemming; and indexing. To identify the best design for fatwa documents, similarity measures were applied to the documents: in the Title, Content and Combination (of both Title and Content). Results reveal that the best structure for Malay language fatwa documents is Title coupled with Cosine similarity.

To Cite This Article: S. Fahmy, I, Nazri, I. Evi, M. Rohana., Design of an Information Retrieval System for Malay Language Fatwa Documents. ***Aust. J. Basic & Appl. Sci.,*** *8(4): 213-218, 2014*

INTRODUCTION

Malaysia, an Islamic country, highly guards the sanctity of Islamic teachings and controls the dissemination of Islamic materials. Fatwas are legal pronouncement in Islam and issued by the Fatwa Committee of the National Council for Islamic Religious Affairs on matters referred to it by the Conference of Rulers.

e-Fatwa and Fatwa Management System are two examples of on-line systems that provide retrieval of Malay language fatwa documents based on user's query. Both systems utilize keyword searches but results are significantly limited due to the terms used and the absence of weighing mechanism. In addition, the semi-structured nature of the database does not support complex queries. Although storing and querying structured data are well understood, there is still no agreement in managing semi-structured data (Heuer and Priebe, 2000).

It is evident that a more efficient system to support the retrieval of Malay Language fatwa documents is needed. The system must not only be able to interpret the content of information from a collection but also rank them according to the degree of relevance based on user's query. This process of interpretation should involves extracting syntactic and semantic information from the documents. As such, this paper aims to propose a design for an information retrieval system for Malay language fatwa documents based on different structures.

2. Information Retrieval:

Information Retrieval is the "representation, storage, organization of and access to information items" (Baeza-Yates and Ribeiro-Neito, 1999). It aims to provide easy access to information in a timely manner. Natural Language is usually used in information retrieval query where text is not always well structured and could be semantically ambiguous (Baeza-Yates and Ribeiro-Neito, 1999).

Corresponding Author: S. Fahmy, Department of Computer Science, Faculty of Computer, Media & Technology Management, TATI University College, 24000 Terengganu, Malaysia.
E-mail: fahmy@tatiuc.edu.my

Information Retrieval Models:

Information Retrieval models can be categorized into two dimensions, *Mathematical Basis* and *Properties of the Model*(Froelich, 2013). The focus of this paper is the mathematical basis of the model namely *Standard Boolean, Probabilistic* and*Vector Space*(Davis *et al*., 1989; Mathieson, 1991; Moore, 1987; Taylor and Todd, 1995).

The*Standard Boolean Model* supports Boolean expression for example queries that accommodate AND, OR, NOT. Although this model works best in viewing documents as a set of words, the major drawback is the difficulty of constructing effective Boolean queries. In addition, this model lacks ranking and weighing mechanism, leading to irrelevant documents being returned.

The *Probabilistic Model* is based on the *Probability Ranking Principle*, which states that a document should be ranked based on the probability of relevance to the query. However, this model requires prior knowledge; lacks structure to represent important linguistic features; and does not maintain Boolean relationships.

Vector Space is an algebraic model for representing text documents as vectors of identifiers, such as index terms. This model is used in information filtering, information retrieval, indexing and relevancy rankings. These attributes make *Vector Space* a suitable model for Malay language fatwa documents.Discussions from this point forward will only focus on the *Vector SpaceModel* due to its relevance in this study as opposed to *Standard Boolean* and *Probabilistic*.

Vector Space Model:

Vector Space Model involves the creation of an index and lexical scanning to identify significant words. Morphological analysis is carried out to reduce different forms to common '*stems*' and computing their occurrences. Ranking is supported by considering the more similar a document vector is to a query vector, the more likely the document is relevant to that query. Key elements in this model are discussed below:

Stemming Algorithm:

Stemming generally means the removal of suffixes in a word. In information retrieval, the collection of documents is described by words in the title (and possibly in the abstract). As such, regardless of the actual location of the word, a document is said to be represented by a vector of words, or *terms*. Terms with a common stem will usually have similar meanings, for example the Malay words PANDANG, PANDANGNYA and PANDANGAN are similar to PANDANG. Stemming reduces the total number of terms in the system, reducing size and complexity of the collection. Some of the stemmer algorithms found in literature include *Nice*(Yang *et al*., 2011), *Text*(Fox and Fox, 2002) and *Porter Stemmers*(Porter, 1980).

Stopwords:

Stopwords are common words in a language such as ITU, BAGAIMANA(Malay) and THE, IT, WHICH (English). There is nodefinite list of stopwords and any group of words can be chosen as the stopwords for a given purpose.It is a practice in designing retrieval system to discard or filter stopwords during indexing. Not only will this eliminate non-significant words that will influence the retrieval process, but also reduces the size of the document up to 50% (Rijsbergen, 1979).

Data Model:

Data model is an abstraction of data, describing how it is represented and used. Data model is constructed based on the documents to be retrieved and has two accepted meaning, *data model theory* and *data model instance*(Grabczewski, 2013). Data model theory is a formal description of how data can be structured and used while data model instance is the creation of a practical data instance for a particular application. Several techniques are available for designing data model such as *Entity-Relationship Model*, *IEDF*, *Object Role Modeling (ORM)*, *Business Rules Approach* and *RM/T*(Hoberman, 2001).

Data Set:

There are two categories of data, *Structured* and *Semi-Structured*. Structured data is a set of records that carry values based on a pre-defined attributes. Semi-structured data, on the other hand, has some structure but may not be rigid, regular or complete and quite often, does not conform to any fixed schema. In semi-structured data, the information that is normally associated with a schema is contained within the data itself. In contrast, relational databases require a predefined table-oriented schema and data managed by the system must adhere to this schema. Most of the approaches to semi-structured data management are based on query language that traverses a tree-labeled representation. Without a schema, data can only be identified by specifying its position within the collection rather than its structural properties.

Indexing:

The simplest form of document retrieval is linear scanning where the whole document is scanned to match the query. To support flexible matching and ranking in lengthy documents, the use of *index* is required. *Indexing* optimizes speed and performance in finding relevant documents for a search query. Without an index, every document in the collection would be scanned, requiring a considerable time and computing power. Among the techniques found in literature for creating index files are *Block Merge*, *Dynamic* and *Distributed Indexing*(Cleverdon, 1991).

Benchmarking:

Queries are generally less perfect in two ways, the return of irrelevant documents and the exclusion of relevant documents. One approach to measure the effectiveness of the information retrieval process is by conducting a binary assessment. A benchmark is usually used where information needs are expressed in terms as queries. Some of the benchmark data sets found in literature includes *Cranfield*, *TREC*, *Reuters-21578* and *20 Newsgroups Collections*(Kent *et al*., 1995).

Retrieval Efficiency:

Tools for measuring the efficiency of the retrieval process include *Precision* and *Recall Rate*s (Fuller and Zobel, 1998). Precision Rate measures the likeliness of a document to be relevant to the query while Recall Rate measures the retrieval of relevant documents. Others measures found in literature include *F-Measure* (combination of Precision and Recall rates) and *Mean Average Precision*(mean of the average Precision scores for each query).

3. System Design:

This section presents the design of a retrieval system for Malay language fatwa documents. There are four phases namely Data Modeling, Indexing, Inverted Index and Search Engine.

Data Modeling:

Relational Approach is used to modeling fatwa documents. This approach has numerous advantages such as enhanced security, concurrency control, recovery, and parallel scalability (Grabczewski, 2013). Fig. 1. illustrates data modeling for fatwa documents.

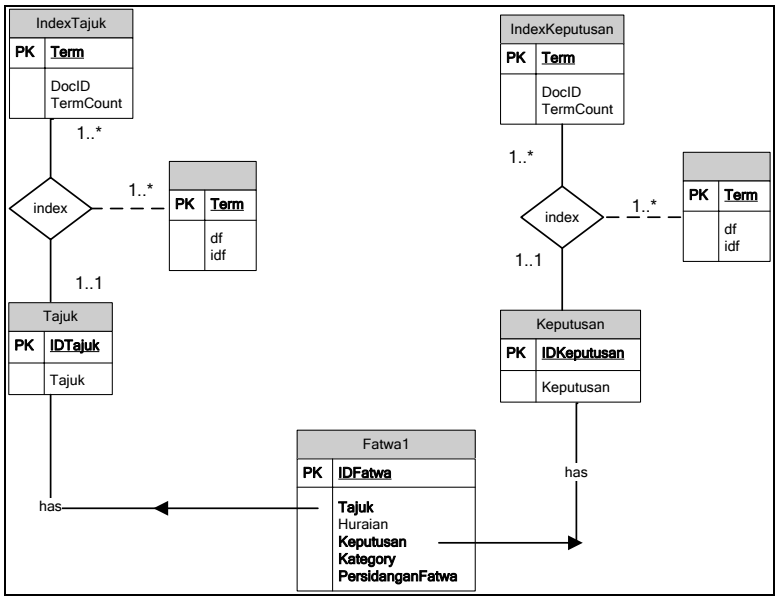

Fig. 1: Data Modeling for Fatwa Documents.

Each field in the fatwa document is indexed separately with its own document frequency (*df*) and inverse document frequency (*idf*) to support zone scoring. The structure of the document is divided into three categories: *Title*, *Result* and *Combination* (of both). *Title* refers to the title of the fatwa while *Result* refers to the content of the document. *Combination* refers to both the *Title* and *Result* of the fatwa document. Each of the categories shares a primary key to enable access in the database (*Title*, *Result*, *Combination*).

Indexing:

In this phase the text are tokenized and linguistic processing is carried out. Tokenizing is a process of slicing character into tokens while linguistic processing is building equivalent classes of the tokens. *Stopwords* such as INI, KEPADA, ADALAH and SEBAGAIMANA were removed from the index. Common Malay stopwords are taken from (M.T. Abdullah and F. Ahmad, 2005). Some of the new stopwords identified in this work are listed in Table 1.

Table 1: New Stopwords for Fatwa Documents.

syarak	tersebut	fatwa
punya	berikut	nas
islam	berlaku	muzakarah

There are cases where a query does not match any tokens in the index due to spelling discrepancy. For instance, the term JAKIM in a query might not match J.A.K.I.M. in the collection. A standard way to overcome this problem is to implicitly create equivalence classes (Manning *et al*., 2007). The stemming process for Malay word is based on an algorithm by Sankupellay and Valliappan (2006).

Inverted File:

Inverted file is created in this phase for timely retrieval of documents based on index. Each occurance of a term is indexed in a dictionary and posting list. Each fatwa document is allocated a unique serial number (*docID*) and the index of this document is a list of normalized terms. The dictionary stores the term, and has a pointer to the posting list for each term. It also stores additional information such as total frequency of the term, and the number of documents in which each term occurs. Each posting list stores the list of documents in which a term occurs and additional information such as term frequency and its location in the document.

Search Engine:

Finally, the search engine of the system is designed using *Free Text Query* and *Similarity Measure*.*Free Text Query*is a free-form input to the retrieval system. A document with a high degree of occurrence is often the best result for the query. A mechanism is used to compute the score for the term, and score for the term and document. To rank the results of the query, the score for the query is computed for each matching document (query/document pair). *A Similarity Measure*is a function which assigns a number to a pair of vectors. Simple similarity measures may count the number of *terms* in the query and document. In this work, three similarity measures are applied to count the number of *terms* in the query namely *Cosine*, *Russel-Rao* and *Dice.*

4. Design Evaluation:

This section presents the experiment that was carried out involving 100 fatwa documents and 16 queries. The aim of the experiment is to evaluate the effectiveness of the retrieval process for different document structures using the chosen similarity measures. The structures are *Title*, *Result* and *Combination* while the similarity measures used are *Cosine*, *Dice* and *Russel-Rao*. *Term Frequency* is used as the weighting scheme.

Cosine is used to calculate the distance between vectors for the items and vectors in the query (Hoberman, 2001) . Fig. 2 illustrates the results of *Cosine* for *Title*, *Result* and *Combination* where the *Result* structure exhibits the best performance.

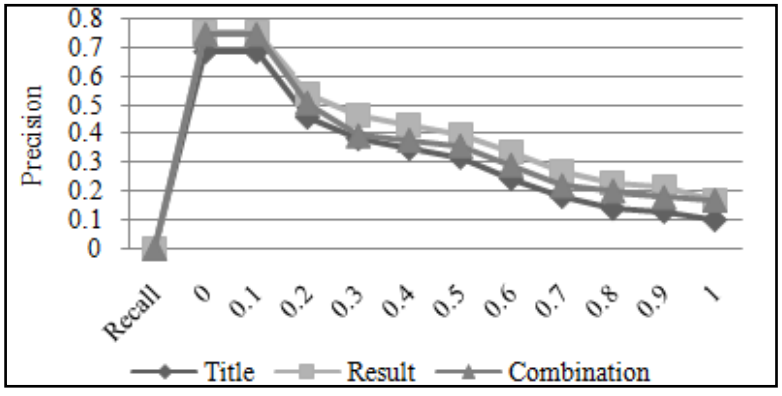

Fig. 2: Results of Cosine Similarity Measure.

Dice similarity measure changes the normalizing factors in the dominator for different characteristic of data (Hoberman, 2001). Using *Dice*, the *Combination* structure exhibits the best performance as illustrated in Fig. 3.

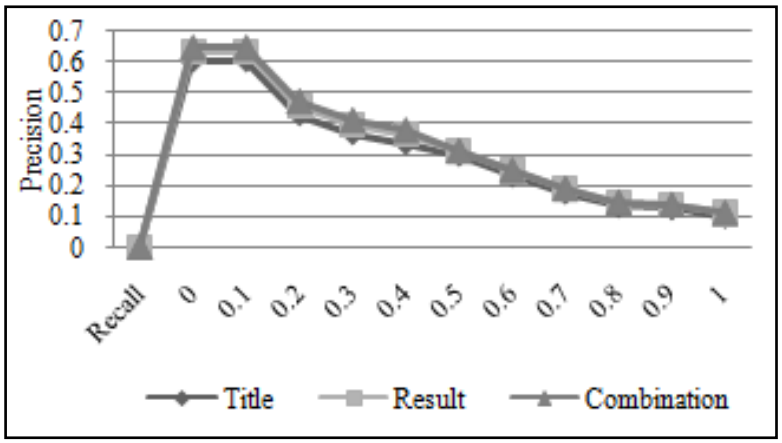

Fig. 3: Results of Dice Similarity Measure.

Russel-Rao normalizes the number of attributes of internal data (Hoberman, 2001). When tested, the *Result* structure exhibits the best performance (Fig. 4).

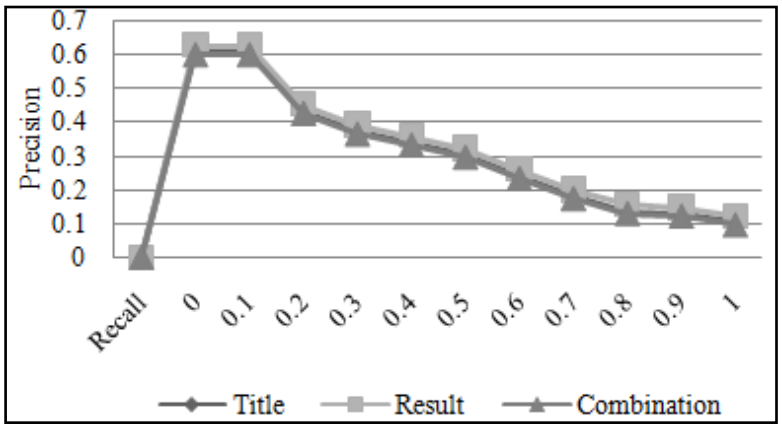

Fig. 4: Results of Russel-Rao Measure.

Based on the results, the similarity measures are then compared by selecting the best performance for each category (Cosine–*Result,* Dice-*Combination* and Russel-Rao–*Document).* Result of this comparison reveals that *Result* exhibits the best performance over the other two similarity measures (Fig. 5).

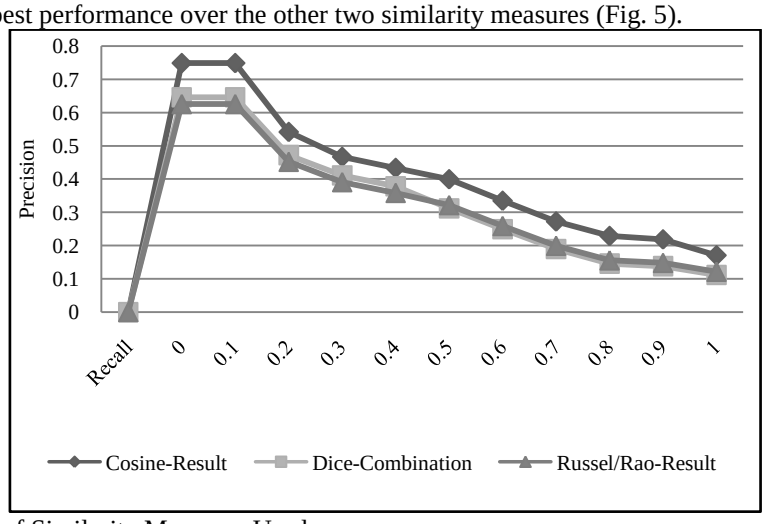

Fig. 5: Comparison of Similarity Measures Used.

It is evident that based on these experiments that Cosine similarity with *Result* structure produces the best results due to the shorter document length than *Combination* as Cosine similarity favors short documents over long documents. *Title* on its own does not yield good results due to the document being too short coupled with limited terms. *Combination*, even lengthy, did not exhibit better results than the other two.

Conclusion and Future Work:

This paper has presented a design for an efficient retrieval system of Malay language fatwa documents. The design of the retrieval system was based on the *Vector Space Model* andevaluated using *Cosine, Dice* and *Russel-Rao* similarity measures. Based on the analysis, *Cosine* exhibits better performance than *Dice* and *Russel-Rao*; and the combination of *Cosine* and *Result* structures proveto be the best approach for fatwa documents retrieval system.

Future works include investigating other similarity measures and weighting schemes. An improvement to the similarity performance of fatwa documents should also include new termssuch as *'zakat harta'* and *'harta sepencarian'*, which is the combination different terms but has a totally different meaning. An improvement to the Malay stemming algorithm is also proposed as a direction for future work.

REFERENCES

Abdullah, M.T. and F. Ahmad, 2005. "Improvement Of Malay Information Retrieval Using Local Stop Words", International Advanced Technology Congress.

Baeza-Yates, R. and B. Ribeiro-Neito, 1999. "Modern Information Retrieval", Addison Wesley: England.

Cleverdon, C., 1991. "The Significance Of The Cranfield Tests on Index Languages", Proceedings of SIGIR, pp: 3-12.

Davis, F.D., R.P. Bagozzi and P.R. Warshaw, 1989. "User Acceptance of Computer Technology: Comparison of Two Theoretical Models", Management Science 01/1989; 35(8): 982-1003.

Fox, C. and B. Fox, 2002. "Efficient Stemmer Generation Project", Conference on AI: Advances in Artificial Intelligence, pp: 655-663.

Froelich, J., 2013. "Information Retrieval", http://en.wikipedia.org/ Last accessed August 2013.

Fuller, M. and J. Zobel,1998. "Conflation-Based Comparison of Stemming Algorithms", Proceedings of the Third Australian Document Computing Symposium Sydney, Australia.

Grabczewski, E., 2013. "Data Model", http://en.wikipedia.org/Last accessed August 2013.

Heuer, A. and D. Priebe, 2000. "Integrating a Query Language for Structured and Semi-Structured Data and Information Retrieval Techniques", University of Rostock Germany, pp: 703-707.

Hoberman, S., 2001. "Data Modeler's Workbench: Tools and Techniques for Analysis and Design", New York: John Wiley & Sons.

Kent, A., M.B. Madeline, U.L. Fred and J.W. Perry, 1995. "Machine Literature Searching VIII", Operational Criteria For Designing Information Retrieval Systems. American Documentation, 6(2).

Manning, C.D., P. Raghavan and H. Schütze, 2007. "An Introduction to Information Retrieval", http://www.informationretrieval.org/Last accessed.

Mathieson, K., 1991. "Predicting User Intentions: Comparing the Technology Acceptance Model with the Theory of Planned Behavior", Information Systems Research, 2(3): 173-191.

Moore, G.C., 1987. "End User Computing And Office Automation: A Diffusion Of Innovations Perspective", Infor, 25(3): 214-35.

Porter, M.F., 1980. "An Algorithm for Suffix Stripping. Program. Automated Library and Information Systems", 14(3): 130-137.

Rijsbergen, C.J., 1979. "Information Retrieval", 2ndEdition.London: Butterworths.

Sankupellay, M. and S. Valliappan, 2006. "Malay-Language Stemmer", University Of Malaya Sunway Academic Journal, 3: 147-153.

Taylor, S. and P.A. Todd, 1995. "Understanding Information Technology Usage: A Test Of Competing Models", Information Systems Research, 6(2): 144-174.

Yang, K., D. Song, W. Jeoung and R. Tang, 2011. "Nice Stemmer", http://ils.unc.edu/iris/irisnstem.htm/Last accessed March 2011.