Electrical Engineering and Informatics
Institut Teknologi Bandung, Indonesia June 17-19, 2007

Terms Visualization for Malay Translated Quran Documents

Normaly Kamal Ismail[1*], Nurazzah Abd Rahman[1], Zainab Abu Bakar[1],
Tengku Muhammad Tengku Sembok[2]

[1] *Faculty of Information Technology and Quantitative Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor Darul Ehsan, Malaysia*
[2] *Faculty of Technology and Information Sciences, Universiti Kebangsaan Malaysia, Bangi, Selangor Darul Ehsan, Malaysia*

A web-based visualization system is developed to visualize the similarity between root words in Malay translated Quran Documents. The visualization of terms used is based on their similarities measures using Cosine and Dice coefficients. The degree of similarity of a term (A) with a processed term (B), can be easily determined by observing the location of term A from term B. The term that is located closer to the processed term is considered more similar to the processed term compared to other terms that are located farther away. The terms created in the processing of one term can be used as potential queries to search relevant documents in the Malay translated Quran manually or electronically. The flexibility of this system to visualize different terms will be discussed. The development of the system involves two stages. The first stage is the processing of 6236 documents of the Malay translated Quran to create a database of all terms. The second stage is the creation of the web-based system using the preprocessing data created in the first stage. Both stages are explored. The visualization will help to improve the understanding of relationship between the terms in this specific domain.

1. Introduction

Thesaurus can be used in information retrieval during query process to help in retrieving relevant documents. Thesauri may exist as synonyms, antonyms or terms co-occurrence (8). The terms co-occurrence thesauri can be generated automatically with no human intervention (8). Synonyms and antonyms thesaurus do not have any value attached to them. On other hand, for terms co-occurrence t, each term will has a value that indicate the degree of association between them. Measurement of association between two terms can be calculated using cosine similarity or dice similarity formula. Visualization of term together with it best terms co-occurrence related to similarities is one of the ways to ensure that good term can be selected as an alternative query in searching for relevant documents. The documents collection has direct influence to the creation of terms co-occurrence. Malay translated Quran documents are used as the test collection in this research. This will contribute to better understanding of the Quran. A visualization of terms co-occurrences of one term will help to determine a membership of the term in any broader terms or topics.

A few approaches have been used to produce a well scattered of terms co-occurrence in the visualization process. The use of two similarities formula in measurements, combination of related and in sequence documents as a documents collection, and scaling of the dissimilarity values are the approaches being used. A web-based platform is used in this system to ensure that this system can be reached by anybody and also can be used together with any other applications in the web. Many functions available in the Internet browsers such as hyperlink and I/O functions help to simplify the execution of the system.

* normaly@tmsk.uitm.edu.my

2. Measurements of Association

A distance measure or a measure of similarity or dissimilarity is needed to measure the degree of association between two terms. There are a number of similarity measures available. The most suitable similarity measures for comparing terms vectors because of their simplicity and normalization are cosine similarity and dice similarity (7).

Let $f(t_i d_k)$ is the frequency of term i in document k. The cosine similarity C is:

$$C(t_i, t_j) = \frac{\sum_N^{k=1} f(t_i, d_k) * f(t_j, d_k)}{\sqrt{\sum_N^{k=1} f(t_i, d_k)^2} * \sqrt{\sum_N^{k=1} f(t_j, d_k)^2}} \quad (1)$$

The dice similarity D is

$$D(t_i, t_j) = \frac{2 * \sum_N^{k=1} f(t_i, d_k) * f(t_j, d_k)}{\sum_N^{k=1} f(t_i, d_k)^2 + \sum_N^{k=1} f(t_j, d_k)^2} \quad (2)$$

where N is the number of unique terms in the collection (6).

The normalization makes the value of $C(t_i,t_j)$ and $D(t_i,t_j)$ lie between zero and one. Value one means that term t_i, and term t_j appear in the same documents at all times and value zero means that both terms never appear in the same documents at all times. Similarity matrices $Sim(t_i, t_j)$ of size N X N for both cosine and dice can be created using Eqs. (1) and (2). Dissimilarity matrix is $Dis(ti,tj) = 0$ implies that the distance between both terms is nil. Hence, both terms are similar. The smaller the value indicates both terms are closer and relevant to each other. $Dist(ti ,tj) = 1$ implies that there is no common appearance in any documents for both terms. Hence, as in Eqn. (3), both terms are not relevant to each other.

ISBN 978-979-16338-0-2

Dissimilarity matrix is:

$$Dis(t_i, t_j) = 1 - Sim(t_i, t_j) \quad (3)$$

for all values of i and j.

Dissimilarity matrices for both cosine and dice can be created and saved in files for further processing in the system development stage. The process and functions to create dissimilarity matrix files in a pre-processing stage are shown in Fig. 1.

3. Experimental Details

The main objective of this research is to develop a web-based visualization system of terms co-occurrences which can realize new resources and in turn can be used in analyzing a specific documents collection.

3.1 Test Collections

The Malay translated Quran (3) will be used in this research. The Quran is arranged in three synchronize ways. Firstly, the Quran is divided into 30 almost equally size of sections. Secondly, the Quran is divided into 114 unequal size of chapters. And finally, the Quran has 6236 sentences called ayah. We called it as ayah collection. One ayah is considered as one document in this research. All the chapters are created from combinations of ayah. All of the chapters have their own unique names. The ayah is recognised by using the name of the chapter where it belongs and the location number of the ayah in the surah. The smallest number of terms in any of the ayah is one and there are many ayat that have the number of terms less than ten. Too many documents with small number of terms will produce many terms co-occurrence with small values and this will effect the visualization of the terms co-occurrence (1). To solve the problem, based on the suggestion by Islamic scholars (3), we combine all related and in sequence ayat to become one document. This compilation reduced the number of documents from 6234 documents to 799 documents and the average number of terms in this compilation is more than the ayat. We call this new compilation as topics collection. We will use both collections as part of the Quran test collection in this research. On top of the two collections, the Quran test collection also consists of stopword list, morphological rules, and dictionary of Malay roots word. Stopword list, morphological rules, and dictionary of Malay roots word are needed in stemming algorithm during the process of creating the dissimilarity matrix.

3.2 Stopword List

A stop word list is a list of high frequency function words considered to have no indexing value, used to eliminate potential indexing terms. These stop words are poor discriminators and cannot possibly be used to identify the document content. An additional of 36 stop words are added in the original list created by (4) which consists of 314 stop words, totaling to 350 stop words. A few examples of the stop words are '*telah*', ' *bagaimana*', ' *selalu*', '*itu*', '*dia*', '*kamu*' and '*yang*'.

3.3 Morphological Rules

Stemming algorithms are used to identify morphological variants and are language dependent. A stemming algorithm for Malay Language written by (4) called Rules-Application Order (RAO) is used in these experiments. To stem Malay text effectively, not only suffixes but also prefixes and infixes must be removed in proper order (4)(9)(1).

3.4 Dictionary of Malay root words

The dictionary being used here is the dictionary of root words, which is developed for the Malay stemming algorithm by (4). This dictionary is later used by (1) and (2) to stem Malay words in evaluating the effectiveness of conflation methods in retrieving Malay documents. This dictionary is used by (5) to stem Malay words in evaluating the effectiveness of clustering techniques in retrieving Malay documents.

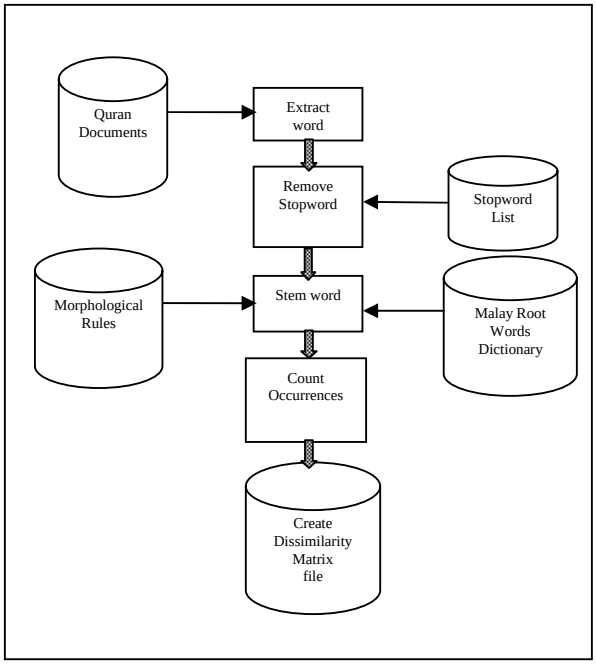

Fig. 1. Pre-processing to Create Dissimilarity Matrix Files

4. System Developments

A web page on which the user can make selections and also can enter a term (root word) is created. The selections are for a similarity formula, a documents collection and a scaling function. The input term is the term that we want to display its terms co-occurrence.

After making the selection, entering an input and pushing the go button, one web page will appear displaying 49 terms co-occurrence and an input term. The input term will appear in the centre of the web page. All terms display on the web page are hyperlinked. When any of the terms is selected and executed, another web page with the same kind of output will appear but now using the selected term as an input term. This web page will maintain the same selections data from the previous execution.

ISBN 978-979-16338-0-2

Scaling of the dissimilarity values are done to ensure that the visualization of the terms co-occurrence are in well scattered manner. There are possibilities that most of dissimilarity values are small and have almost the same values. If this occur, all the terms co-occurrence will appear close to the border of the web page and will not produce a meaningful layout as in Fig. 2.

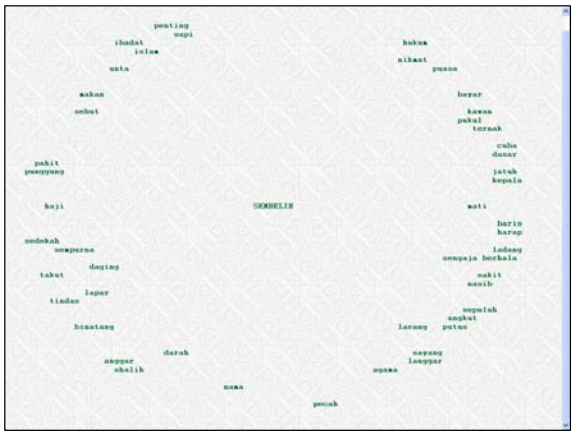

Fig. 2. Terms Co-occurrence on Ayah Collection using Dice Coefficient

4.1 Web Page Creation

The process of locating terms in the web page is as follows. A 2-dimensional display with 50 rows (y-axis) and 110 columns (x-axis) on the common one screen page of the Internet browser will be used. Texts overlaps are not allowed in the common Internet web page. To avoid congestion, only 50 terms will be used and displayed. Dissimilarity values of all the terms co-occurrence (associate to the input term) that we are going to display were produced in the pre-processing stage. These values are ready to be retrieved and to be used in this stage. The values are sorted in a descending order because in this way we can select the top 49 of the similarity values. The transformation of similarity values (range between 0.0 and 1.0) to the range of 0 to 25 (half of the number of rows used) have to be done. Scaling of the x (column) and y (row) values need to be done because of the different size of the row and the column in the web page.

The input term is placed at the centre on the web page which is at the location (55,25) in the x and y axis coordinates. We can freely place the terms co-occurrence anywhere surrounding the input term in the x and y axis. Theorem Pythagoras is used to calculate the distance between the input term and the terms co-occurrence.

Let z (the transform dissimilarity value) is the distance between the input term and one of the terms co-occurrence. Then, we can randomly choose the value of x which is x <= z. Using Theorem Pythagoras as in Eqn. (4), we can find the value of y.

$$y = \sqrt{z^2 - x^2} \qquad (4)$$

where $x \geq 0$ and $y \geq 0$.

The sign of x and y values are determined randomly. Then, the real coordinate of the term co-occurrence is determined by adding the value of x and the value of y with 55 and 25 (x and y axis location of the input term) respectively. If an overlap occurs, the location will be recalculated for a new available location.

In a scaling situation, the scaling factor must be chose correctly to ensure that the scaled dissimilarity values are more than 0.0. Only the non-zero dissimilarity values are scaled. To solve this problem, the scaling factor is chose base on the smallest dissimilarity value for the input term.

5. Results, Discussion and Conclusion

In Fig. 2, the dice formula, ayah collection, no scaling function are chose and the term "*sembelih*" is entered. The output display shows that almost all terms co-occurrence appear far away from the input term "*sembelih*". The terms co-occurrence produce a circle-like shape which is close to the border of the web page. However, from the display of the terms co-occurrence, we can make a conclusion that the term "*sembelih*" belongs to "*haiwan ternakan*" class.

When the scaling function is used to the previous output display (Fig. 2), several terms co-occurrence which are not close to the input term are now become closer to the input term such as "*larang*", "*daging*", "*lapar*" and "*sengaja*" as in Fig. 3.

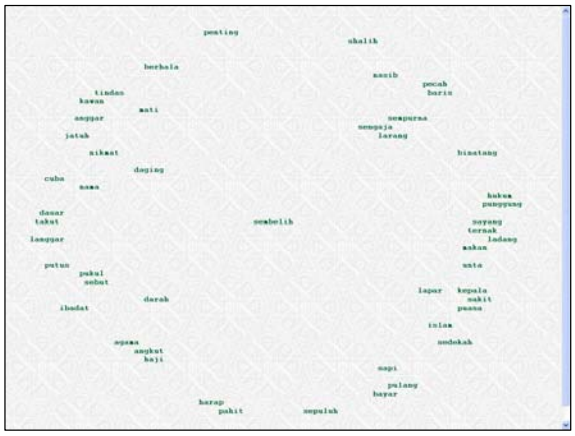

Fig. 3. Terms Co-occurrence on Ayah Collection using Scaling Function

When the ayah collection is replaced by the topics collection, several new terms co-occurrence appear as in Fig. 4. As overall, all the terms co-occurrence are closer to the input term if we compare to the previous output display (Fig. 3). The number of terms co-occurrence that are close to the input term are also increased.

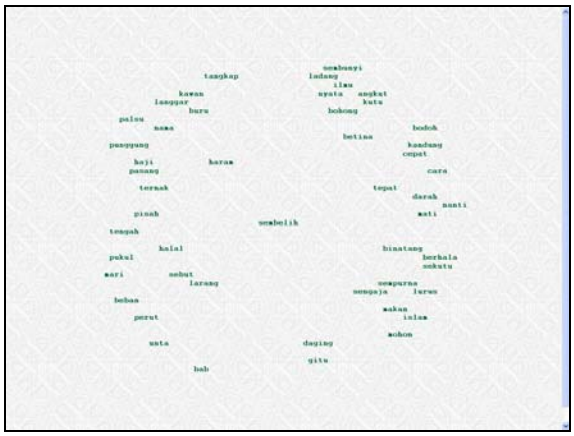

Fig. 4. Terms Co-occurrence on Topics Collection using Dice Coefficient

When we replace the dice formula with the cosine formula, changes are discovered where several terms co-occurrence that are close to the input term are changing place as in Fig. 5. The changes give new meanings to several terms co-occurrence with associate to the input term.

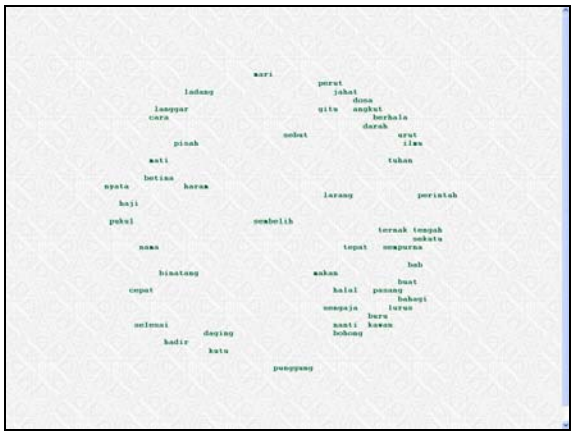

Fig. 5. Terms Co-occurrence on Topics Collection using Cosine Formula

The execution of a term co-occurrence "*larang*" from the previous output display (Fig. 5) using hyperlink function that maintains the same selections data and using "*larang*" as an input term produces an output display as in Fig. 6. As we can see that several terms co-occurrence which appear closest to the input term "sembelih" from the previous output display (Fig. 5) are also appear closest to the input term "*larang*" such as "*tepat*" and "*haram*". From both output displays (Fig. 5 and Fig. 6), we can make a conclusion that the terms "*haram*" and "*tepat*" are the most suitable candidates as alternative queries for "*sembelih*" and "*larang*" to search for relevant documents.

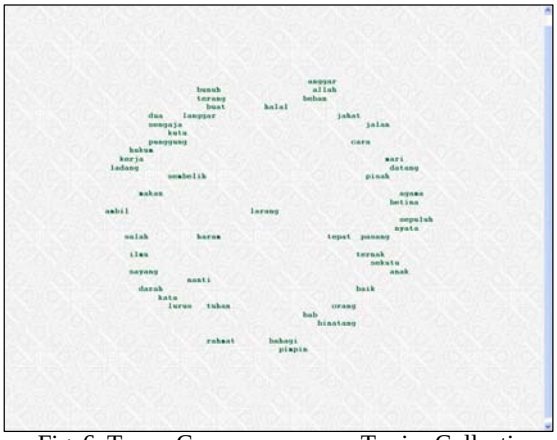

Fig. 6. Terms Co-occurrence on Topics Collection using Hyperlink

As a conclusion, this web-based term visualization system can be used to realize new resources from the selected domain. The new resources in turn can be used in the process of analyzing and understanding the specific domain or other related domains such as a Malay translated hadist documents collection (5).

References

(1) A. B. Zainab: Evaluation Of Retrieval Effectiveness Of Conflation Methods On Malay Documents, *Ph.D. Thesis*, Universiti Kebangsaan Malaysia (1999).

(2) A. B. Zainab A. B. and A.R. Nurazzah: Evaluating the Effectiveness of Thesaurus and Stemming Methods in Retrieving Malay Translated Al-Quran Documents, *Lecture Notes in Computer Science 2911*, Springer-Verlag, Berlin Heidelberg, Germany, pp.653-662 (2003).

(3) Al Quran dan Terjemahan: *Madinah Munawarah*, Saudi Arabia.

(4) A. Fatimah: A Malay Language Document Retrieval System: An Experimental Approach And Analysis. *Ph.D. Thesis.* Universiti Kebangsaan Malaysia (1995).

(5) A.R. Nurazzah, A.B. Zainab, A.B. and I. Normaly Kamal: Experiments On Clustering Techniques In Retrieving Malay Translated Hadith Text Documents, *Proceedings of Brunei International Conference of Engineering and Technology (BICET'05)*, Bandar Sri Begawan, Brunei (2005)

(6) D. Widows: Geometry and Meaning, *CLSI Publication*, Stanford, California (2004).

(7) G. Salton: Automatic Text Processing, *Addison Wesley*, Reading, Mass (1989).

(8) R. Korfhage: Information Storage and Retrieval, *Wiley Computer Publishing*, New York, New York (1997).

(9) T.M.T. Sembok, M. Yussoff and F. Ahmad: A Malay Stemming Algorithm for Information Retrieval, *Proceedings of the 4th International Conference and Exhibition on Multi-lingual Computing*, (1994).