2D TEXT VISUALIZATION FOR THE RETRIEVAL OF MALAY DOCUMENTS

NORMALY KAMAL ISMAIL
Computer Science Department
Universiti Teknologi MARA
40450 Shah Alam, Selangor
MALAYSIA
normaly@tmsk.uitm.edu.my

TENGKU MOHD TENGKU SEMBOK
International Islamic University Malaysia
P.O. Box 10
50728 Kuala Lumpur
MALAYSIA
tmts@ftsm.ukm.my

Abstract: - Search engine applications like Google and Yahoo present their results in the form of one-dimensional linear list that usually comprise three times of the screen size per page and several number of pages. The results are displayed in the list of inconsistent declining ranks without displaying its rank values. The one-dimensional linear list display of the results data will cause classification of the results data meaningless. New queries relating to the original query are available, but its relationship strength values are not provided An application that can display all the result data in a two-dimensional text visualization within one page and circular form is proposed. The relationship strength of the result data with the query can be evaluated by finding the distance between the location of the result data to the center of the circle. Classifications that are made in the form of text and color can easily apply to the application. Malay translated Al-Quran and Malay translated hadith are used as corpuses for the application. Three functions in the application display the relationship between words and words, between words and documents, and between documents and documents. Various combinations of formulas can be used to find the values of these relationships that will be used as the rank values in the application. This, two-dimensional text visualization (TDTV), application is evaluated using two mechanisms. First, by solving a task and then, follow by answering the usability questionnaire. The results from the task section show that the variety of related documents can be retrieved in a reasonable time frame. The results from the usability questionnaire show about 75 percent of the respondents agree that the two-dimensional text visualization (TDTV) application is better than applications that display its results in one-dimensional linear list.

Key-Words: - Information retrieval, visualization, classification, web-based, usability.

1 Introduction

Malay language is widely used in countries such as Malaysia, Indonesia, Brunei, Singapore, Thailand and the Philippines. More than 250 million people use it, especially in Indonesia and Malaysia. Although there is some difference in the number of Malay words from one country to another country, it can still be understood. More than 90 percent of those using the Malay language is Muslims. Thus, the corpus of Malay language is the most influential in this society is the Malay translated Quran [1] and the Malay translated Hadith [2][3]. However, the online search engines especially for the Malay language which uses both corpuses are almost nonexistent. Several search engines for malay language like [4] and [5] using exact matching techniques that are known to have many shortcomings. There is one search engine on the Internet, for the Malay translated Hadith, uses vector techniques to display results according to the Hadith relevancy [6]. This search engine uses the same approach with other popular search engines, such as Google and Yahoo, to display its search results in one-dimensional linear list. One-dimensional linear list display of the search results

ISBN: 978-1-61804-064-0

has several disadvantages and will be described in the next section.

2 Problem Statements

Information retrieval engine used in the Internet browser will display all the search results in the form of one-dimensional linear list. Three main weaknesses can be detected when a one-dimensional linear list is used in the display of the search results. The first problem is when search results are too many; the one-dimensional linear list display will take more than one page. Scroll function has to be used in order to view the last part of the search results that are located in the end part of the linear list. If the results are to be preserved in one page, the search results will be broken down into several pages accessible through a hyperlink. However, additional activities such as using the mouse function to perform the hyperlink have to be done and will take more time. The second problem is the documents obtained and displayed in one-dimensional linear list method do not include their ranked. They only list according to the ranked where the highest rank is placed at the beginning of the display and the lowest rank is placed at the end of the display. This causes users deceived by the position of the documents in this one-dimensional linear list display. This is because the ranked values do not decrease evenly. This might be expected by the users when they examine the uniform one-dimensional linear list display. The third problem is the difficulty of implementing the various classifications when the one-dimensional linear list of search results is used. While there is classification using names of the verses for the Quran or using hadith narrators name for the hadith, but it is difficult to make any conclusions from it because the one-dimensional linear list of search results of the Quran and the hadith will extend over several pages. Applying several classifications in the one-dimensional linear list of the search results may produce very compact display and difficult to understand.

3 Problem Solutions

Solution to the problems described earlier on information retrieval engine by developing an application that displays search results in a circular form as the Fig.1 that displays the query results "ekor".

Fig.1 Display search results in a circular form.

A total of forty documents are displayed in one page with their position calculated from the center of the circle. The closer the documents to the center of the circle, the stronger their relationship with the query "ekor" are. Thus, the users can see the differences relationship strength, whether they are strong or weak based on the distance. Also, clearly visible on display are the documents classifications. For the Malay translated Hadith, two classifications are formed that are the classification based on the narrators of the Hadith and the classification based on the topics for the Hadith. Hadith topics classification are done using texts, and Hadith narrators' classification are done using two different colours. For the Malay translated Quran, two classifications are constructed. The first classification is based on the name of place where the verse was revealed. Two colours are used. The second classifications based on the name of the verse and texts are used. The problem of information retrieval engines for Malay documents that they only focus on the relationship between queries and relevant documents only. No attempt to provide guidance in the formulation of queries that are associated with the query being used. No efforts to highlight other documents that may be associated with documents that relate to the query. Therefore, the solution is to construct three functions that implement three different applications and then combine them. This application will apply two-dimensional text visualization (TDTV) system. An example of application that display the output results for a word (query) and the documents relationship are in Fig.1 above.

ISBN: 978-1-61804-064-0

4 Application Development

The development of TDTV application consists of two phases that are the pre-processing phase and the application construction phase. In the pre-processing phase, several mathematical formulas are used and the values produced will be implemented and used in the construction phase of the application.

4.1 Pre-Processing Phase

Searching relevant documents are derived from a word or group of words called phrase. So every word or phrase must have a value that can be linked to each document. For the initial stage, the relationship values between each word with all documents are calculated first. Then, the relationship values between each phrase with all documents will be generated from the combination of the relationship values of each word in the phrase with all documents. To improve the retrieval of the related documents, words that regard based on its co-occurrence to the query word/phrase in the documents can be used as an alternative word for the query in the searching process. A table contains the relationship values based on co-occurrences for every word with all the words in the documents should be built. Again, the relationship values of each phrase with all the words can be produced by combining all the co-occurrences values of the words in the phrase with all the words. Finally, the similarity of a document with other documents can also be used to retrieve the relevant documents. Therefore, a table contains the relationship value between each document with all documents should be built.

Three types of tables can be formed for the three relationships before. The first table is the table of values that form the relationship between each word with all documents in a corpus and is called the word-to-document matrix. The second table is the table of values that form the relationship between each word with all words in a corpus and is called word-to-word matrix. The third table is the table of values that form the relationship between each document with all documents in a corpus and is called the document-to-document matrix. To illustrate these relationships, values that represent these relationships are calculated using the vector space formula. Vector space is the point in the space that represents any value produced by various combinations that come from several sources. This allows us to use the space more than three dimensions without us to illustrate. Several similarity formulas are used as to construct the matrices such as Cosine similarity, Dice similarity and Jaccard similarity formulas. All calculations can be with or without a weight associated with them. The tf-idf (term frequency – inverse document frequency) formula is used to calculate the weight. Two corpuses are used that are the Malay translated Qur'an and the Malay translated hadith. Each corpus requires all together 14 tables of relationship values matrix as in table 2. Formulas found in [7] and [8] and calculation examples of the relationship values can be found in [9]. Stemming technique by Fatimah used in words processing and can be found at [10] and [11]. However, in actual applications, only one table of relationship matrix used in each relationship. The formula that produces the best distribution for the display will be selected. However, this selection will be discussed in a future publication.

4.2 Application Construction Phase

Once the databases are built, then the construction of these applications can be implemented. The application uses Internet browser and bases on two-dimensional display with the x-axis and y-axis. The application will retrieve and display documents or words that make x and y coordinates as its position. Position of each document/word on the display is based on the value generated in the relationship matrices discussed earlier. Three displays will be built using three types of relationships that have been built in the form of matrix tables. The first display is a display of the word/phrase with words most associated with it in-term of its co-occurrence. The values from the word-to-word relationship matrix are used in this display. The values of the word-to-document relationship matrix are used to create a second display. It displays the query word /phrase together with the documents that are relevant to the query word/phrase. The final display shows a document as query and documents which relate to the query. The values from the document-to-document relationship matrix form the position of the documents in the display. Since the two-dimensional coordinates approach (x-axis and y-axis) is used as the display area, coordinates (0, 0) be the center of the display. In this position, (0.0), a document or a word used as a query is located. Words or documents associated with the word or document used as a query would appear around the center of these coordinates (0.0). The position of the words or documents with the center of the display depends on the values taken from the relationships mentioned before. The matrices values range is between 0 and 1. Value 0 means

no relationship at all, while the value of one indicates the strongest relationship. For the document-to-document relationship matrix, the value of 1 means they are two documents that have the same contents. For the display positions, the smaller value means the short distance is between one another on the display. The short distance between one document to the document used as a query means the relationship between them is strong. Therefore, each value in the matrix will be converted to the actual distance values as follows:

Object distance to the center of the display
= 1 - the relationship value (1)

As a convenience, the background of the display prints certain numbers by using bright colours that do not interfere with the actual output, and it is, in fact, the values of the x-axis and y-axis that can be used as a reference. Values selected for x-axis is from the range of -50 to 50, and the y-axis is from the range of -28 to 28 which will cover nearly a one-page display. Position on the y-axis is not the actual position of a word or document because this value was reduced to 55% of actual value to overcome the larger line/row size than the size of the column. For wide computer displays, the 70% of the actual values are used, so the actual distance can be achieved. The new values will be used to locate the documents on the display. The same method also used for the two relationship matrices that are word-to-word and document-to-document relationships. By knowing the distance of two objects, say a single word with another word, also with one position is located at the center coordinate (0.0), then the other position is easily determined using the Pythagorean Theorem, provided that the object can be located at any position in the display either on the right, left, above or below of the center of the display. The values of x and y coordinates of the object's position should be sought. Both of these values must be in the range of 0 and 1. Firstly, the distance will be changed from the range of 0 to 1 to the range of 0 to 50 by multiplying the value to 50. Number 50 was chosen because 100 columns (x-axis coordinates) for the display.

$$\text{distance} = \text{distance} \times 50 \tag{2}$$

For the rows on the display, only 58 rows are used. The original value of the y coordinates will be reduced to 55% and will be done after the coordinate x is calculated first. For the y coordinate, its value can be found at random, and it must be less than the distance of the original as follows:

$$y = \frac{\text{distance}}{1000} \times \text{random}(1-999) \tag{3}$$

Once the value of y is obtained, the value of x coordinate can be calculated using the *Pythagorean* Theorem.

$$x = \sqrt{\text{distance}^2 - y^2} \tag{4}$$

Then the value of y coordinate reduced to 55% from its initial value as follows:

$$y = \frac{y \times 55}{100} \tag{5}$$

The x and y coordinates are rounded to the nearest whole number. The rounding process and the process of reducing the value of y coordinate will result in the value calculated is not 100% accurate. This small error does not significantly affect the display that was built later. After that, both, the x and y coordinates will be determined at random either they are positive or negative value. These x and y coordinates values will be the location of the document/word. Fig.2 shows the difference position of the retrieved documents in the output display. If all the retrieved documents have the same distance to the center of the display, a circle will be formed in the display as Fig.3.

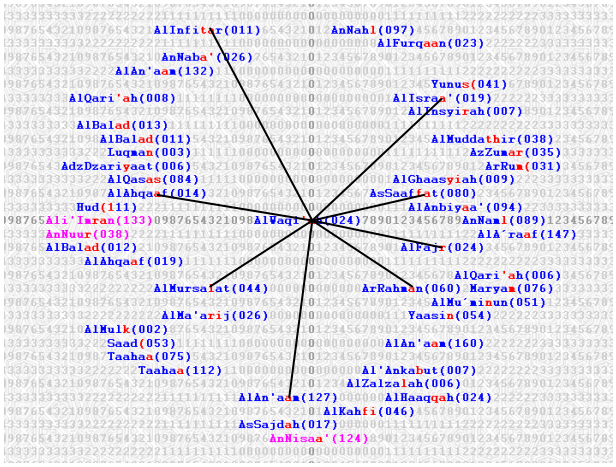

Fig.2 Various distances of documents to the center of the display.

This document/word will be located in either at the top or bottom or left or right of the center of the display (0, 0). Objects placed in the position of x

and y coordinates are not one character length, but using several characters. So, the maximum length has to be set.

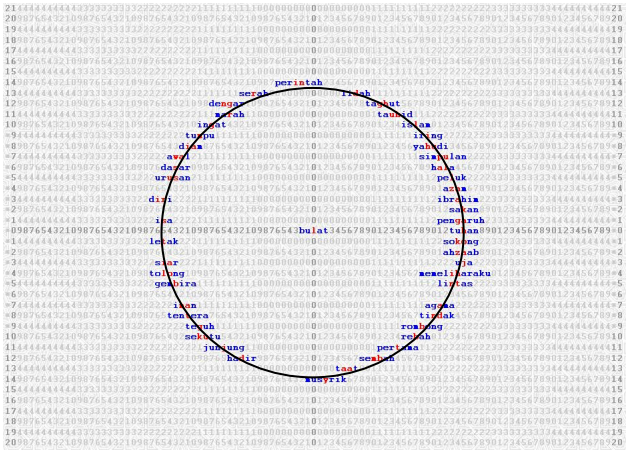

Fig.3 Documents with the same distance to the center of the display

Thus, the overlap would occur. If the position of a document/word overlaps with the already existing document/word (as calculated earlier), the same calculation process should be done repeatedly until the non-overlapping new value of x and y coordinates found. A never-ending loop may occur if all the appropriate positions on the display have been used, causing the program halt and no output display will be produced. Improvements implemented by creating classification of documents displayed using texts and colours as in Fig.4. The classifications are performed manually.

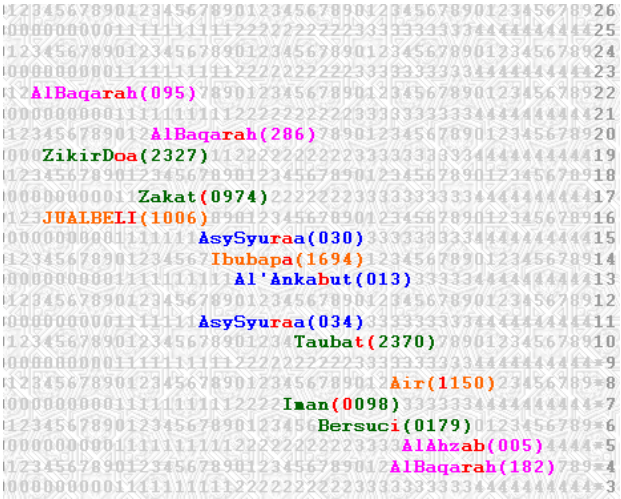

Fig.4 Classifications using texts and colours.

5 Application Evaluation.

Application evaluation is performed in two ways that are to do a task and to answer the usability questionnaires for the applications. Fourteen respondents perform the task and answer the questionnaire. The purpose of the task is to measure the effectiveness of the application in terms of the retrieval performance and the time taken to retrieve the documents. The two applications uses in the task are the applications that display its results in one-dimensionnel linear list and the application that was proposed (TDTV). All respondents are given a same query, and they should retrieve twelve documents from the two applications and from the two corpuses mentioned before. The time is taken for each execution of the application. Documents obtained will be analysed by an expert in the Quran and Hadith to determine its relevancy. The results are as Table 1.

Usability questionnaire will be answered by the respondents only after they complete executing the given task. A total of 56 questions is in the usability questionnaire and questions are all addressed to the three displays in the proposed application and the proposed application as a whole.

Table 1 Results of the searching activities by the respondents.

Relevancy	Number of al-Quran verses retrieved	
	Application VTDD	Application OSLSD
Not Relevance	1	0
Relevance	35	27
Ave. Time	6.5 sec.	4.4 sec.

Table 2 Results for the usability questionnaires in percentage.

Application	Answer Scale (Percent)				
	Strongly not agree	Not agree	Not sure	Agree	Strongly agree
Word ⇔ Word	0	0	44%	44%	12%
Word ⇔ Document	0	0	12%	63%	25%
Document ⇔ Document	0	0	12%	63%	25%
Overall	0	0	25%	63%	12%

Respondents should be answered using the following five scales: 1 - Strongly agree 2 - disagree, 3 - not sure, 4 - agree, and 5-strongly agree. The questions in the usability questionnaire are about the text size, the distance between the text

ISBN: 978-1-61804-064-0

and the center of the display, the display size, the colors used and its combination, the display background, the types of queries, the execution speed, the errors recovery, the learning time and the performance comparison with other applications. The results are as Table 2.

6 Conclusion

Two-dimensional text visualization application is successfully developed and implemented. The concept of displaying the retrieved Malay documents and Malay words in the form of two-dimensional successfully applied in this application. Display results in the circular form is implemented. Three functions successful implemented. These functions are functions that perform and display the word-to-word, word-to-document and document-to-document relationship matrices. Values that form the matrices were also produced. In addition, a number of classifications of the results successfully applied to this application.

Users acceptance is evaluated by carrying out tasks and answer to the usability questionnaire for the application. Acceptance percentage of the respondents to this application is 75%. Even for a question that asks whether this application helps to find the desired documents, almost all respondents agreed that this application helps to get the documents sought. Therefore, the application of two-dimensional text visualization not only successfully developed and is due to function as required, but accepted by users as the better application to access information on the source of Islam in the Malay language.

References:

[1] Sheikh Abdullah Basmeih. *Tafsir pimpinan Ar-Rahman kepada pengertian al-Qur'an*. Kuala Lumpur: Bahagian Hal Ehwal Islam, 1980.

[2] Daud, M. *Terjemah Hadis Shahih Muslim*, Volume I-IV, Darel Fajr Publishing House, Singapore, 2003.

[3] "e-Hadith", June 6, 2011, <http://ii.islam.gov.my/hadith/hadith.asp>.

[4] Hamidy, Z. , Thaha,N., Fachruddin, HS., Ariffin, J., & Zainuddin AR. *Al-Imam Al-Bukhary, Terjemahan Hadis Shahih Bukhari* Volume I – IV, Darel Fajr Publishing House, Singapore, 2002.

[5] "Mutiara Hadis", June 6, 2011, <http://sigir.uitm.edu.my/webhadis/>.

[6] "Berita Harian Online", June 6, 2011, <http://www.bharian.com.my/>.

[7] Korfhage, R. R., *Information Storage and Retrieval,* Wiley Computer Publishing, New York, New York, 1997.

[8] Salton, G. *Automatic Text Processing, Addison Wesley*, Reading, Mass. 1989.

[9] Widdows, D. *Geometry And Meaning,* CSLI Publishing, Stanford University, 2004.

[10] Sembok, T.M.T, Yussoff, M. & Ahmad, F., A Malay Stemming Algorithm for Information Retrieval. *Proceedings of the 4th International Conference and Exhibition on Multi-lingual Computing*: 1994. 5.1.2.1-5.1.2.10

[11] Fatimah Ahmad. *A Malay Language Document Retrieval System An Experimental Approach And Analysis*. Ph.D. Thesis. Universiti Kebangsaan Malaysia, 1995.