Semantic Method for Query Translation

Mohd Amin Mohd Yunus, Roziati Zainuddin, and Noorhidawati Abdullah
Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Abstract: *Cross language information retrieval (CLIR) presents huge ambiguous results as polysemy problems. Therefore, the semantic approach comes to solve the polysemy problem which that the same word may have different meanings according to the context of sentences. This paper presents semantic technique on queries for retrieving more relevant results in CLIR) that concentrates on the Arabic, Malay or English query(s) translation (a dictionary based method) to retrieve documents according to query(s) translation. Therefore, semantic ontology significantly improves and expands the single query itself with more synonym and related words. The query however is to retrieve relevant documents across language boundaries. Therefore, this study is conducted with the purposes to investigate English-Malay-Arabic query(s) translation approach and vice versa against keywords and querywords based on total retrieve and relevant. Keywords and querywords retrieval are evaluated in the experiments in terms of precision and recall. In order to produce more significant results, semantic technique is therefore applied to improve the performance of CLIR.*

Keywords: *Semantic ontology, semantic query, CLIR, dictionary-based.*

Received September 2, 2010; accepted May 24, 2011; published online March 1, 2012

1. Introduction

Generally query affects on retrieval results in terms of its accuracy and effectiveness to those users information seeking processes. A lot of researches have been done upon on the requirement to accurate results. One of the considerations is semantic approach in order to be applied on Cross Language Information Retrieval (CLIR) [3]. Therefore, a lot of search engine have this approach to get the significant results as domain and ontology [4]. Domain however depends on how it is developed to contribute the better retrieval results [5]. Thus it is beneficial to improve retrieval performance to meet the necessary results for the respective user inquiry [9, 10]. In addition, the retrieval process is most important part to the user query for achieving outstanding retrieval results [14]. The use of semantic technique which is applied on query and documents can be successfully to retrieve more relevant documents as a whole result.

The query is then translated into another language by translation dictionary. This translated query is helpful for the performance of the retrieval system. In order to conduct a research, data are gathered accordingly and respectively as a collection of documents. This translation dictionary is useful for translating the query to meet those words in documents to be worth result. An additional dictionary is developed and used for removing the unnecessary words in the query as known as stop words. It removes the meaningless words in the query before continuing to annotate each word in the query for better relevant judgment of information results.

In order to annotate each word in the translated query in another language, it is important to prepare the semantic dictionary to link each word to others to have annotated words for each word in the translated query. It leads the relevant annotated words to each word in the translated query. Hence a lot of relevant words annotation can explain retrieved information in the result based on step by step in the processes of semantic translated query.

2. Related Works

There are several studies conducted to prove the significance of semantic method to be applied to various fields and aspects such as semantic web and information [3, 5]. Therefore, the semantic approach is more significant to be integrated with queries which have been done for better related information even though between languages [1, 2]. The results more relevant and required documents retrieved or displayed. Thus Yang [21], explains Distributed Semantic Indexing (DSI) addresses both the data quality and search performance. With the ability of summarizing content information and guiding data distribution, the proposed approach is distinguished firstly logic-based representation and concise abstraction of the semantic contents of multimedia data, which are further integrated to form a general overview of a multimedia data repository. Secondly application of linguistic relationships to construct a hierarchical metadata based on the content signatures allowing imprecise queries and finally achieving the optimal performance in terms of search cost.

However, Rinaldi [17] proposes a novel metric to measure the semantic relatedness between words.

Our approach is based on ontology represented using a general knowledge base for dynamically building a semantic network. This network is based on linguistic properties and it is combined with our metric to create a measure of semantic relatedness. O'Hara and Wiebe [13], describe on how semantic role resources can be exploited for preposition disambiguation. The main resources include the semantic role annotations provided by the Penn Treebank and FrameNet tagged corpora. The resources also, include the assertions contained in the Factotum knowledge base, as well as information from Cyc and conceptual graphs. a common inventory is derived from these in support of definition analysis, which is the motivation for this work. M`arquez *et al.* [11], assess weaknesses in semantic role labeling and identify important challenges facing the field. Overall, the opportunities and the potential for useful further research in semantic role labeling are considerable.

3. Introduction to Basis Mathematical Approach

Let's word (*W*) as total words which consists of word 1 (*w1*), word 2 (*w2*), word 3 (*w3*) and the rest words (wn) in the search field as shown in equation 1. Then, every word (*w*) has its own semantic words (*Z* or *S* or *T*) as many as possible which are available in the semantic data or dictionary (*Y*) as shown in equation 2 and 4. *T* represents the total semantic words derived from a single query which consists of a few words (*W*) in the search fields.

$$W = \sum_{i=1}^{n} i \tag{1}$$

$$T = (\sum_{w=1}^{W} w \in \sum_{t=1}^{Z} t) \subseteq Y \tag{2}$$

where n is the last number of word and *i* is the first word. Therefore user can input words as many as they want as long as the total of retrieval results from the input words is not influenced. In addition, those words again contribute their own semantic words where one word can yield few words which depend on the words provided in the semantic dictionary. Regarding the result of the words, let *D* as total retrieval documents related to each word if in each document for the first word (w∩d1, w∩d2, w1∩d3,….w1∩d6236), followed by (w2∩d1, w2∩d2, w2∩d3,….w2∩d6236) and last word should be (wn∩d1, wn∩d2, wn∩d3,….wn∩d6236) with the additional results of first synonym or semantic word (t1∩d1, t1∩d2, t1∩d3,….t1∩d6236), followed by (t2∩d1, t2∩d2, t2∩d3,….t2∩d6236) and last word should be (tn∩d1, tn∩d2, tn∩d3,….tn∩d6236). Those relevant ayats (verses) are counted as shown in equation 3:

$$D = \sum_{d=1}^{6236} W \cap d + \sum_{d=1}^{6236} Z \cap d \tag{3}$$

The meaning of 6236 is related to the total ayats or documents or verses or files in whole quran as a collection. Semantic (S) documents are the results to be derived from multiple Ds for retrieving results which is depending on each word (W) that consists of related synonym words (Z or S or T) come from semantic dictionary (Y).

$$S = \sum_{i=1}^{n} W1i + \sum_{i=1}^{n} W2i + \sum_{i=1}^{n} W3i ... + ... \sum_{i=1}^{n} Wmi \tag{4}$$

Figure 1 shows that the expanded query which base on a single query. It starts with source language of a single query (Q) to be translated into target queries (TQ1, TQ2, TQ3, … TQn) of different languages. Each word (w1, w2, w3, … wn) carries with a few of semantic or synonym words (t1, t2, t3, … tn) according to the related words in the semantic data or dictionary. Those collective words which consist of original and semantic words are matched to those words in all ayats (verses or documents) to retrieve relevant multilingual verses. Those relevant potential collective semantic and synonym words in the expanded query are matched with relevant results.

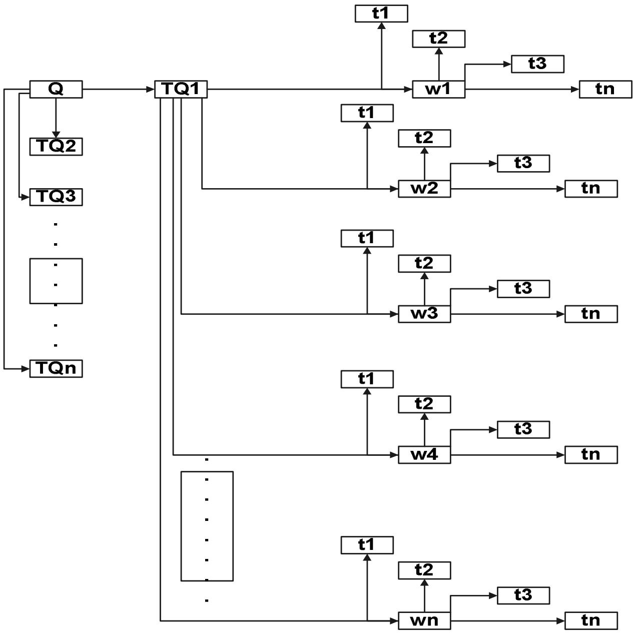


Figure 1. The flowchart of expanding an single query.

4. Experimental Approach

Qur'an documents are the scope in these experiments which are preparing the original and holly Quran in classical Arabic language in text, Malay Quranic documents translation collection [8] which is used by Zainab and Nurazzah [27] as a domain in their research and English Quranic documents translation collection [12]. Each collection has 114 surahs and 6236 documents for every language which three languages get involved. Every document has numbers beginning with q which means query and followed by numbers with first three numbers denote chapter and last three number denote verse for every filename such as q034006 but in the filename the number of surah begins with "." or dot, followed by the number of surah and the number of ayat begin with "," or comma followed by the number of ayat as for example ".34,6". All documents are as flat files in UTF-8, ASCII or EBCDIC text and searching process is through pattern matching [6].

Testing Malay query words are taken from Fatimah's [7] collection as natural language queries and the English as well as Arabic query words are translated from the Malay query words. Fatimah has obtained them by looking at few procedure put forward by Popovic and Willet [15] as well as Salton [19]. Each query would be separated and broken into keywords and replaced by target language. For example, if query is Malay, so it is called as source language and the target language is English or Arabic. Thus English as an example represents the translated word to retrieve English documents and if the query is English, the target language is Malay or Arabic. The dictionary lists 1325 Arabic, Malay and English words in different flat files as well as 36 Malay query words selected. The translation refers to the same index between Malay natural query languages [7] and the translation of the natural query language in English. When the keyword is Malay, then reference is to the English word at the same index or when the keyword is English and then reference is to the Malay word at the same index. It is considering word by word in the text files.

The overview of the process begins with the query, matching, retrieving, evaluation and retrieval result. The query term can be Arabic, Malay or English and also, keywords or querywords according to words in the query. The query can be viewed by two choices which are keyword and queryword. If the query is keywords, the results retrieved according to word by word results and redundant document names existed rankly. But querywords, retrieved according to the whole words as one at all and only when no redundant or unique document names retrieved if merged. Query translation can replace the origin query in to another language of the query. This translation is most important for those languages to investigate those information retrieval results. All documents are saved in ".txt" format file for UTF-8, ASCII or EBCDIC text. For searching process, word by word matching is used in the process. The matching words refer to the words similarity between query and documents in retrieving process. The query submitted to the system is also, represented by translated query that is used to search the related files. Translated query is processed by removing the meaningless words or stopwords. Then, query can be translated into another language if needed by the dictionary.

The query results consist of word by word result or keyword or the whole words and full phrase or querywords. The relevant documents retrieval comes from al-Quran multilingual documents collection which is from Arabic, Malay and English collection. For this research, when the query is Malay words, the words are translated into English or Arabic vice versa. Figure 2 shows the flowchart of retrieving semantic results based on a single query to be optimized and expanded for its semantic words. In order to get translated semantic words, the words in a single query are translated and matched depending on the indexed words which are available for words translation and their semantic simultaneously in semantic data after stop words have been removed based on stop words data. Then, original words and their semantic words are collected to search those documents which are related to the expanded query. All relevant multilingual documents are displayed for evaluation of each query of each language.

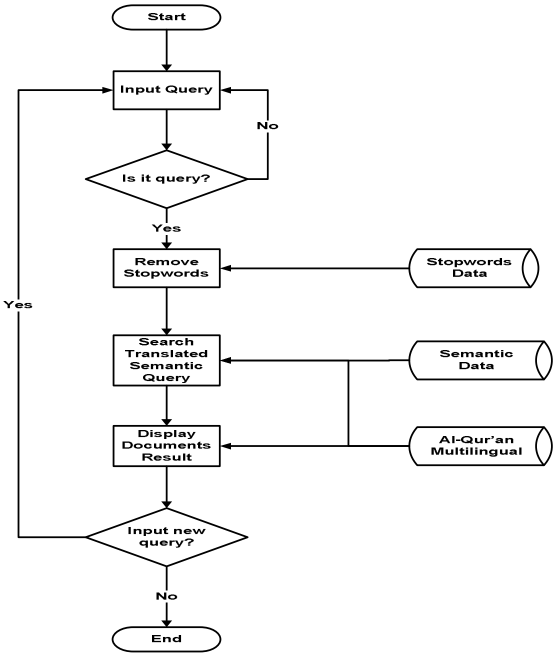


Figure 2. The Flowchart of retrieving semantic results based on single query.

5. Result

All results of the query(ies) translation are referred to the natural language queries of Malay [7] and then translated into Arabic and English [22, 23, 24, 25, 26] queries in this study. Every query is tested to evaluate each result which is matched with manual result as total relevant documents for respective query [7]. There are six queries for each language in Table 1 for testing the results performance of Semantic and Non-Semantic whereby each query is tested for its Semantic and Non-Semantic results. Hence, the evaluation technique is used for recall results whereby equation 5 shows the formula to calculate the percentage of recall [18]. Expanded query is explained based on Figures 1 and 2 to retrieve potential results in order to meet the higher recall percentage of each query.

$$Recall(\%) = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents} * 100 \qquad (5)$$

Table 1. Natural language queries.

No	Malay	English	Arabic
1	membuktikan kewujudan alam	prove the existence of natural	اثبات الوجود السريالي
2	tanggungjawab anak ibu bapanya	responsibility of parents child	مسؤولية الطفل الأم والد
3	kelebihan berpuasa sembahyang kesihatan kewajipan	the merit of fasting prayer health obligations	ميزة الصيام الصلاة الصحة التزام
4	mewajibkan sembahyang jumaat	require prayer friday	متطلبات صلاة الجمعة
5	kisah lelaki tertidur gua beratus-ratus tahun	story of men sleeping caves hundreds of years	قصة اهل الكهف
6	tuntutan berperang jalan allah	claim struggle in the way of allah	المطالبة محاربة طريق الله

Figure 3 shows the percentage against recall for semantic Arabic/Malay Query English Document. The Recall K (word by word option) indicates the highest total queries and percentage 91-100 %. It means that the total relevant and better results meet the user's query accurately compared to others. It is also, followed by Recall Q which means query 1 and 2 have 100%. Meanwhile, recall K has the highest total of relevant judgment and explain the document location come from. Both of them have a lot of retrieval results with relevant results to the user's query. Non-semantic results are quite lower in recall percentage than semantic results.

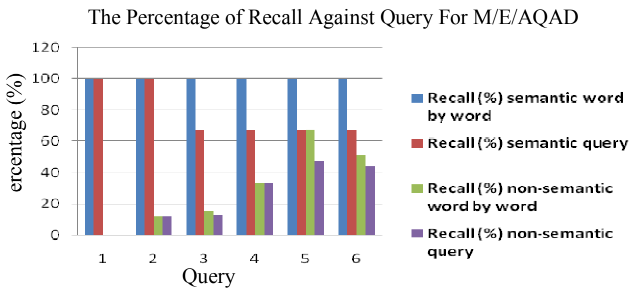


Figure 3. The recall percentage for semantic Arabic/malay query english document.

Figure 4 shows the percentage against recall for semantic Malay/English Query Arabic Document. The Recall K (word by word option) indicates the highest total queries and percentage 91-100 % for query 1, 2, 4, 5 and 6. It means that the total relevant and better results meet the user's query accurately compared to others. It is also, followed by Recall Q which means query 4 and 5 have 100%. Meanwhile, recall K has the highest total of relevant judgment and explain the document location come from. Both of them have a lot of retrieval results with relevant results to the user's query. Non-semantic results are also, having quite lower in recall percentage than semantic results.

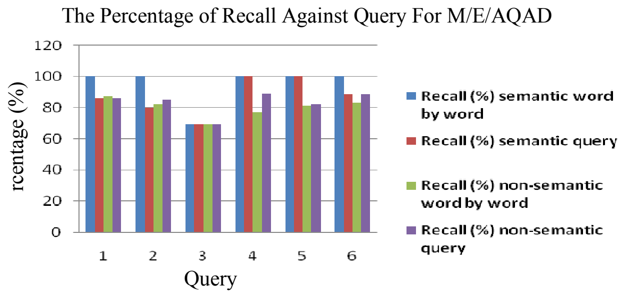


Figure 4. The recall percentage for semantic malay/English query Arabic document.

Figure 5 shows the percentage against recall for semantic Arabic/English Query Malay Document. The Recall K (word by word option) indicates the highest total queries and percentage 91-100 % for query 2, 4 and 6. It means that the total relevant and better results meet the user's query accurately compared to others. It is also, followed by Recall Q which means query 2 has 100%. Meanwhile, recall K has the highest total of relevant judgment and explain the document location come from. Both of them have a lot of retrieval results with relevant results to the user's query. It shows that non-semantic results are quite lower in recall percentage than semantic results.

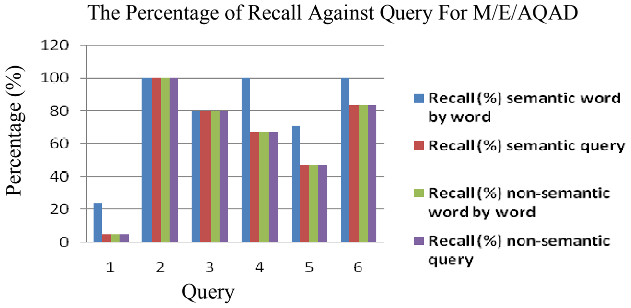


Figure 5. The recall percentage for semantic Arabic/English query malay document.

Table 2 explains average percentages of recall and precision of semantic experiments. The highest percentage of average recall for K and Q are Malay/English/Arabic Query English Document which carries 100.00% and Malay/English/Arabic Query Arabic Document at 87.27% respectively. It means that retrieval results yield a lot of documents with good effectiveness. These semantic results can be more significantly improved compared to [16] which carry approximately 70% of recall. The comparison is related to the performance evaluation result E-Te and E-Ta experiments which retrieved 86% of recall [20]. [16, 20] conducted the experiments without using semantic technique which can improve the effectiveness of CLIR especially on their experiments. Therefore, non-semantic results are quite lower in average of recall percentage than semantic results.

Table 2. Average percentages of recall of semantic query experiments.

No	Experiments	Average of Semantic Recall (%)		Average of Non-Semantic Recall (%)	
		K	Q	K	Q
1	MQMD/ EQMD/ AQMD	78.89	63.50	63.50	63.50
2	EQED/ MQED/ AQED	100.00	77.78	29.88	24.98
3	AQAD/ MQAD/ AQAD	94.87	87.27	79.87	83.27
4	HTO- E[16]	70			
5	E-Te & Te-E [20]	86			

6. Discussion

Empirical experiments are conducted with the purposes to investigate semantic results according to semantic query. Thus semantic query consists of meaningful and synonym but different words. Those words however depend on provided words in dictionary or file. Furthermore, it is also, conducted to investigate the performance between keywords and querywords based on total retrieve and relevant for each retrieval process. The retrieval however, included the unnecessary documents because of the translation polysemy. This research also, is being applied in retrieving Quran documents collection which consists of holly classic Arabic language and then followed by English and Malay translated documents with queries compared to single query without synonym words in searching retrieval. It leads more and more relevant results displayed.

The results are shown for each of them according to languages which are Arabic, English and Malay. Their effectiveness and retrieval and relevant percentage are included as part of evaluation and analysis. Keywords (K) and querywords (Q) are different processes of retrieving results. K usually has relevant information about documents retrieval results which means word by word option. Table 2 shows that the experiments results in terms of average of recall percentages. Translation of query is not affecting on recall. Experiments on English [22, 23, 24, 25, 26] and Arabic documents show the better average Recall for K and Q respectively.

7. Conclusions

Semantic query show better performance for retrieving results in K than Q in the specific language. The queries need a semantic and synonym data in dictionary to help and translate words and then expand those words for their synonym words in target language for queries or documents. Thus translated queries with collective synonym word can retrieve quite relevant and related documents to meet closely 100% user's documents requirements. Effective semantic approach depends on how many synonyms words provided in the dictionary which is unlimited references for each word in the query. It means that the more synonym words provided for each word in the query to be processed, the more relevant and related results to be retrieved at higher recall percentage.

Acknowledgements

This research has been funded by the University of Malaya, under the grant number (PS210/2009B) and this work has been done and fully sponsored by University of Malaya PhD scholarship.

References

[1] Abdelkader D., "Ontology-Based Intelligent Mobile Search Oriented to Global e-Commerce," *The International Arab Journal of Information Technology*, vol. 7, no. 1, pp. 28-34, 2010.

[2] Ali G. and Maseud R., "An Ontology-Based Semantic Extraction Approach for B2C eCommerce," *The International Arab Journal of Information Technology*, vol. 8, no. 2, pp. 136-170, 2011.

[3] Aufaure M., LeGrand B., Soto M., and Bennacer N., *Metadata-and Ontology-Based Semantic Web Mining in Web Semantics and Ontology*, Idea Group Publishing, 2006.

[4] Baazaoui-Zghal H., Aufaure M., and Soussi R., "Towards an On-Line Semantic Information Retrieval System Based on Fuzzy Ontologies," *Journal of Digital Information Management*, vol. 6, no. 5, pp. 375-385, 2008.

[5] Baazaoui-Zghal H., Aufaure M., and Mustapha N., "A Model-Driven Approach of Ontological Components for On-line Semantic Web Information Retrieval," *Journal on Web Engineering*, vol. 6, no. 4, pp. 309-336, 2007.

[6] Elly J., "The Study of Existing Malay Algorithm Performed on Words Beginning with 'D'," *BSc Thesis*, Universiti Teknologi MARA, 2000.

[7] Fatimah A., "A Malay Language Document Retrieval System, An Experiment Approach and Analysis," *Thesis Ijazah Doktor Falsafah*, Universiti Kebangsaan Malaysia, 1995.

[8] Hamidy H. and Fachruddin H., "Tafsir Quran," *Translation*, Klang Book Centre, 1987.

[9] Lopez V., Motta E., and Uren V., "PowerAqua: Fishing the Semantic Web," *in Proceedings of the European Semantic Web Conference*, 2006.

[10] Lopez V., Uren V., Motta E., and Pasin M., "AquaLog: An Ontology-Driven Question Answering System for Organizational Semantic Intranets," *Journal of Web Semantics*, vol. 5, no. 2, pp. 72-105, 2007.

[11] M`arquez L., Carreras X., Litkowski K., and Stevenson S., "Semantic Role Labeling,"

Computational Linguistics, vol. 34, no. 2, pp. 145-159, 2008.

[12] Muhammad T. and Muhammad M., *Interpretation of the Meaning of the Noble Quran*, Dar-us-Salam Publications, 1999.

[13] O'Hara T. and Wiebe J., "Exploiting Resources for Preposition Disambiguation," *Computational Linguistics*, vol. 35, no. 2, pp. 1-34, 2008.

[14] Pasi G., "Flexible Information Retrieval: Some Research Trends," *Mathware and Soft Computing*, vol. 9, no. 1, pp. 107-121, 2002.

[15] Popovic M. and Willett P., "The Effectiveness of Stemming For Natural-Language Access to Slovene Textual Data," *Journal of the American Society For Information Science*, vol. 43, no. 5, pp. 384-390, 1992.

[16] Prasad P., Tune K., and Varma V., "Improving Recall for Hindi, Telugu, Oromo to English CLIR," *in Proceedings of Advances in Multilingual and Multimodal Information Retrieval*, *Lecture Notes in Computer Science*, vol. 5152, pp. 103-110, 2008.

[17] Rinaldi M., "An Ontology-Driven Approach for Semantic Information Retrieval on the Web," *ACM Transaction*, *Internet Technology*, vol. 9, no. 3, 2009.

[18] Salton G. and Mcgill J., *Introduction to Modern Information Retrieval*, Mcgraw-Hill, New York 1983.

[19] Salton G., "Experiments in Automatic Thesaurus Construction for Information Retrieval," *in Proceedings of Ifip Congress*, pp. 43-49, 1971.

[20] Sujatha P., Dhavachelvan P., and Narasimhulu V., "Evaluation of English-Telugu and English-Tamil Cross Language Information Retrieval System using Dictionary Based Query Translation Method," *International Journal of Computer Science and Information Security*, vol. 8, no. 2, pp. 314-319, 2010.

[21] Yang B., "DSI: A Model for Distributed Multimedia Semantic Indexing and Content Integration," *ACM Transaction, Multimedia Computing Communications Application*, vol. 6, no.1, 2010.

[22] Yunus M., "Short Query Translation: A Dictionary-Based Approach to Cross Language Information Retrieval, Master of Computer Science," *MSc Thesis*, Universiti Teknologi MARA, Malaysia, 2008.

[23] Yunus M., Zainuddin R., and Abdullah N., "Semantic Query for Quran Documents Results," *in Proceedings of IEEE Conference on Open Systems*, Malaysia, pp. 1-5, 2010.

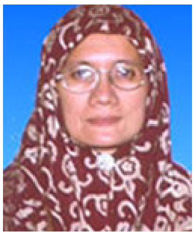
[24] Yunus M., Zainuddin R., and Abdullah N., "Semantic Speech Query via Stemmer for Quran Documents Results," *in Proceedings of International Conference on Electronic Devices, Systems and Applications*, Malaysia, pp. 17-21, 2011.

[25] Yunus M., Zainuddin R., and Abdullah N., "Visualizing Quran Documents Results by Stemming Semantic Speech Query," *in Proceedings of International Conference on User Science and Engineering*, Malaysia, pp. 209-213, 2010.

[26] Yunus M., Zainuddin R., and Abdullah N., "Semantic Query with Stemmer for Quran Documents Results," *in Proceedings of IEEE Conference on Open Systems*, Malaysia, pp. 40-44, 2010.

[27] Zainab A. and Nurazzah A., "Evaluating The Effectiveness of Thesaurus and Stemming Methods in Retrieving Malay Translated Al-Quran Documents," *in Proceedings of the 6th International Conference on Asian Digital Libraries*, Springer-Verlag, vol. 2911 pp. 653-662, 2003,

Mohd Amin Mohd Yunus is currently studying a PhD research related to speech recognition, semantic information and translation, and information visualization at University of Malaya, Malaysia. He has qualifications from MARA University of Technology, Malaysia. His studies started with diploma in computer science in 2002, BSc in Information System Engineering in 2004 and Master Msc degree of science in computer science in 2008. He was an IT lecturer within three years started on 2005 until 2008 at Islamic



Roziati Zainuddin is a professor in the Department of Artificial Intelligence, Faculty of Computer Science & Information Technology, University of Malaya, Malaysia. Her current research interests include visualization of Al-Qur an, development of software for Malay text-to-speech synthesis and speech recognition, HMM-Based speech synthesis and recognition for Malay and Arabic languages, computational fluid dynamics, e-learning, fractal analysis, speech processing, scientific multimedia visualization of medical process, computer vision and image processing.

Noorhidawati Abdullah is a senior lecturer in The Department of Information Science, Faculty of Computer Science & Information Technology, University of Malaya, Malaysia. Her current research interests include incorporating digital library in e-learning platform, e-learning for malay manuscripts and database concepts and implementation.