A Robust Framework for Web Information Extraction and Retrieval

Rayner Alfred, Gan Kim Soon, Chin Kim On, and Patricia Anthony

***Abstract*—The large volume of online and offline information that is available today has overwhelmed users' efficiency and effectiveness in processing this information in order to extract relevant information. The exponential growth of the volume of Internet information complicates information access. Thus, it is a very time consuming and complex task for user in accessing relevant information. Information retrieval (IR) is a branch of artificial intelligence that tackles the problem of accessing and retrieving relevant information. The aim of IR is to enable the available data source to be queried for relevant information efficiently and effectively. This paper describes a robust information retrieval framework that can be used to retrieve relevant information. The proposed information retrieval framework is designed to assist users in accessing relevant information effectively and efficiently as it handles queries based on user preferences. Each component and module involved in the proposed framework will be explained in terms of functionality and the processes involved.**

***Index Terms*—Information retrieval, information retrieval framework, semantic web.**

I. Introduction

Information retrieval (IR) is a process that extracts and retrieves information that is relevant to user based on the queries posted. IR deals with many aspects including the representation, storage, organization and retrieving information from data sources. Furthermore, these data sources can be accessed offline or online and they can be categorized into structured, semi-structured or unstructured data. The origin of the IR research can be traced back to ancient times when librarians kept information related to articles or books using catalogue cards [1], [2] and earlier works related to information retrieval can be found in 1950 [3]. The advent of computer has brought the IR system to a new level as computers are capable of processing large volume of data in order to extract and retrieve relevant information [4]. The increase of capacity and computational power has contributed to the rapid growth of unstructured data. For instance, with the advent of World Wide Web(WWW) making the information available online through hyperlink, the research attention of IR have been shifted to Web IR and it is increasingly gaining popularity. Among significant IR tools for WWW IR are the search engines. In order to retrieve information from the WWW, search engines with different capabilities and algorithm shave been developed. However, the advancement of Internet made information available growth exponential through time and a robust framework for web information extraction and retrieval is critically required to process the overloaded unstructured data.

Big data in today's business and technology environment has contributed to the complexity in accessing and retrieving relevant information for decision making [5]. For instances, there are 2.7 Zeta bytes of data exist in the digital universe today. Facebook stores, accesses, and analyzes 30 over Petabytes of user generated data and more than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide. In 2008, Google was processing 20,000 terabytes of data (20 petabytes) a day.

The rapid growth of unstructured data is also creating a lot of problems. YouTube users upload 48 hours of new video every minute of the day and 571 new websites are created every minute of the day. According to Twitter's own research in early 2012, it sees roughly 175 million tweets every day, and has more than 465 million accounts. There are 100 terabytes of data uploaded daily to Facebook and 30 billion pieces of content shared on Facebook every month. The data production will be 44 times greater in 2020 than it was in 2009 [5].

Thus, more advanced search tools besides search engine are required to tackle this massive and overloaded information. Some of the existing search engines include Yahoo[1], Google[2], Bing[3], Alta vista and etc. Besides that, many information retrieval algorithms have been invented in the researches of search engines [6] and most of them are focusing on a generic search in certain topic. There are also many other information retrieval tools developed for retrieving relevant information that are designed to search for online information that includes READWARE [7] and ontology-based information retrieval [8]. More related works will be discussed in the next section. The aim of this paper is to propose a robust IR framework for retrieving information from the web.

This paper is organized as followed. Section II describes some of the works related to information retrieval. Section III describes the general overview of the proposed robust IR framework and also discusses each component in the proposed framework in details. Section IV concludes this

Manuscript received October 30, 2013; revised December 23, 2013. This work has been supported by the Long Term Research Grant Scheme (LRGS) project funded by the Ministry of Higher Education (MoHE), Malaysia under Grants No. LRGS/TD/2011/UiTM/ICT/04.

Rayner Alfred, Gan Kim Soon, and Chin Kim On are with the COESA, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia (e-mail: ralfred@ums.edu.my, g_k_s967@yahoo.com, kimonchin@ums.edu.my).

Patricia Anthony is with the Department of Applied Computing, Faculty of Environment, Society and Design, Lincoln University, Christchurch, New Zealand (e-mail: patricia.anthony@lincoln.ac.nz).

[1] http://www.yahoo.com
[2] http://www.google.com
[3] http://www.bing.com

paper.

II. Related Work

The traditional IR technology focuses more on content analysis, which are mostly used by search engine. The major methods involved are full text scanning, inversion, signature files and clustering [9]. Full text scanning is a straight forward method that is designed to locate string term or substring in the text content [10]. The signature files approaches create string signature by hashing the words and superimposed coding. The separation storage of the signature and original accelerates the searching process because the signature file size is much smaller [11]. Inversion is a method that inverts the selected index keywords that are used to describe the document to increase the retrieval speed [12]. Generally, a vector space representation is used to represent documents in a vector for ease of indexing and retrieval calculation [13]. Clustering related documents will speed up the retrieval process for relevant information [14]. Integrating semantic background into the vector space representation of documents enhances the precision and the relevancy of the retrieved information. Other methods such as natural language processing (NLP) [15], Latent Semantic Indexing (LSI) [16] and neural network [17] can be applied in order to improve the information retrieval systems.

In addition to that, the emerging trend of the semantic web technology had been widely adopted in information retrieval due to its capability to store multiple descriptions about a single object [18]. For instance, a large-scale text analytics called Seeker and an automated semantic tagging of large corpora called SemTag have been developed [19] and this is one of the early adoptions of semantic web technologies in IR. The utilized semantic web such Resource Description Framework (RDF), Resource Description Framework Schema (RDFS) and a taxonomy based disambiguation algorithm (TBD) have been used to successfully annotate 264 millions of web pages automatically by generating approximately 434 million semantic tags. KIM is another semantic platform that is used for information extraction and retrieval which was developed based on GATE [20]. This platform provides a framework for knowledge and information management as well as services such as automatic semantic annotation, indexing and retrieval of documents. KIM adopted the semantic web technology such RDFS and Web Ontology Language (OWL) for semantic annotation and content retrieval based on semantic queries. This platform utilizes RDF(S) repositories, ontology middleware and reasoning in annotation, indexing and retrieval process. Fang Li and Xuangjing Huang have developed an intelligent platform for information retrieval that addresses the information retrieval problem from three aspects that includes providing a domain specific IR for filtering process, providing concept based IR for words ambiguities, and finally, providing a question and answering mechanism to provide answer to user [21]. A java based platform for context retrieval based on probabilistic information retrieval has also been developed and this is called Okapi [22]. The Okapi is designed and implemented based on dual indexes, relevance feedback with blind or machine learning approaches and query expansion with context.

Collaborative IR is another framework that was introduced to retrieve relevant information by reconciling all information from more than one user [23]. This framework involved three main disciplines which include information retrieval, human computer and supported cooperative work. Most of the collaborative IR systems only focus on the two former disciplines. More and more researchers are interested in the field of collaborative IR due to the proliferation of social network hub such as Facebook, LinkedIn and etc. However, there are still rooms for improvement as this domain is quite new compared to the traditional IR systems. Other researches related to information retrieval include the intelligent semantic search framework that focuses on the semantic of the indexed content [24], an ontological knowledge and context based information retrieval for the personalization search [25], multimedia content retrieval systems [26].

III. A Robust Framework for Web Information Extraction and Retrieval

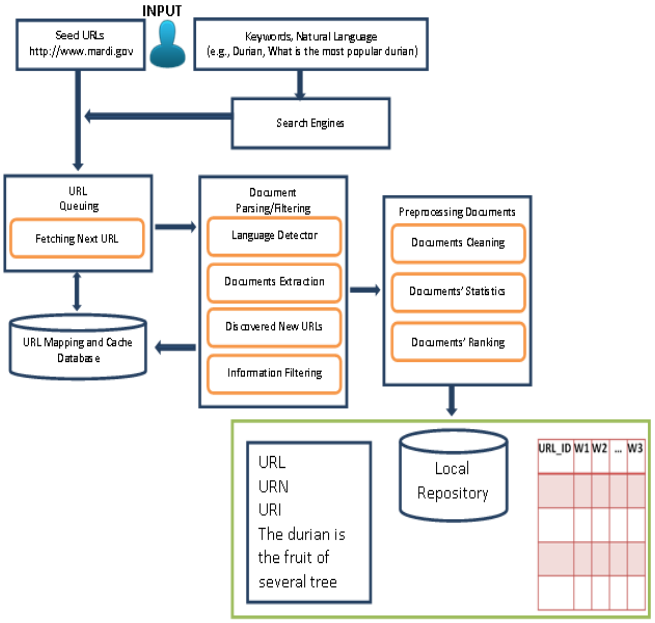

Fig. 1. Information retrieval framework.

In this section, the proposed robust framework for Web Information Extraction and Retrieval will be described in details. Fig. 1 shows the architecture of the information retrieval framework. As depicted in Fig. 1, it is shown in the framework that an URL will be keyed in by the user and this URL will be used as a seed URL for the crawler to start the crawling process in order to search for the input text. If the user does not provide any seed URLs, then the system will invoke any search engines (e.g., Google, Yahoo and Bing) to gather all relevant URLs related to the text input specified by the user. The user will also have the option to specify all URLs that will be excluded in the crawling process and also the number of documents to be extracted and retrieved from the World Wide Web (WWW). User may also limit the depth of the crawling process in order to reduce the crawling process for more relevant documents.

The input(s) from the user that includes the seed URL(s) or

the return results from the search engines will be queued and stored in a database for the retrieval process. The database will keep track the queued URL(s) and update the status of the each queued URL(s) whether it has been crawled before in order to increase the efficiency of the crawling system. Besides that, the database also keeps track the relations between hyperlinks for the construction of the web graph for web indexing and web relationship analysis. Another important role of the database is to keep track of the search history search performed by the user and this data is very important for future enhancement of a recommender system.

The document parsing and filtering module consists of four functions namely language detector, document extraction, discovered new URL(s) and information filtering. In this module, the fetched document obtained from the fetched URL will be scanned by the language detector function in order to determine language used for the content of the web page. This is because the system will need different language processing tasks for different languages. The content of the web site will be extracted by the document extraction function and then processed in order to get a set of newly discovered URL(s). Finally, an information filtering process will be performed to determine whether the content is relevant to the query specified by the user or not. If the contents of the page is relevant, it will be saved and stored for the preprocessing documents module and the newly discovered link with be sent to the database and the queuing URL(s) for the crawling process.

In the preprocessing document module, the extracted document is cleaned in which all the tags (e.g., HTML) will be removed. Then, the documents will be transformed into a vector space representation by using the TF-IDF weighting scheme [27]. Finally, all these documents will be ranked and returned back to the user. User feedback can be considered in enhancing the retrieval process by customizing the filtering algorithm based on users' feedbacks. The text documents and the document representation (e.g., TF-IDF vector representation of the documents) will be stored for future references.

A. Language Detector

The purpose of the language detector function is to determine the language of the web page content. There are several existing methods that can be used to detect the type of language used such as the charset encoding [28], dictionary based approach and n-gram algorithms [29]. This is performed so that documents with different documents can be categorized into the appropriate groups for further processing tasks. In this work, the ASCII encoding and a dictionary-based approach are used to implement the proposed framework. This is because the proposed framework is designed to handle English and Malay documents. If other languages are detected, they will be stored in different directories. Some works related to bilingual information retrieval [30] include the task of expanding the encoding character to Unicode in order to have more documents that can be categorized during the retrieval process.

B. Document Extraction

After the language detector function determined the content language of the web page, the document extraction function will extract the web page content semi-structure or unstructured web page. Wrappers are software tools built from wrapper induction and information extraction concept. This tool can be used to extract information automatically based on a set of extraction rules. These rules can be specified by the user or generated automatically by applying the existing machine learning, pattern marching and other heuristic approaches [31]. Besides that extracting the content, the entities in the content are recognized. For English content, the entities are recognized using existing establish system such GATE [32] and open Calais system [33]. As for Malay content, the entities are recognized using a Malay NER system which developed by our teams members. The entities recognize for the content them will be used as terminology to tag and index the documents for future queries.

C. Discovered Links

The discovered links function will also extract set of hyperlinks that link out from the current web page document. The discovered function is similar to a crawler function where it traversal the HTML page structure to extract the link in this page and send to the URLs queue. The discovered links function will filter out the uncrawled URL links to send to the URLs queue and also the database. However, the extracted URLs will be stored in the URLs queue after the content of the web page relevancy is explored in the information filtering function in the next section.

The framework is not just discovering link in to be crawl, however, link analysis can be carried for analyzing the determined quality of the linked for future query retrieval [34]. There are several methods that can be used in order to perform the link analysis that includes PageRank, Weighted PageRank, HITS algorithms and others [35]. The link analysis can conducted in a context based analysis of the search query which will be store as a separate graph.

D. Information Filtering

The information filtering function will determine whether the document is relevant to the user. Information filtering is the process to help people find the valuable information [36]. Different approaches have been adopted for information filtering such as natural language processing, machine learning, ontology based and others [37]. There are two approach adopted in the information filtering function. If the user is performing a generic search, an adaptive term frequency method will used for the filtering process. This method will use the average term frequency the threshold to determine whether the document is relevant to the user or not. The most common threshold metrics used for the filtering process are mean and standard deviation [38]. If the user performs a specific search, a list of the related terms will be retrieved from the DBpedia ontology. DBpedia and WordNet have been adopted to enhance the performance of the information retrieval task by expanding the query [39], [40].There are two ways to implement the DBpedia ontology is either the download the DBpedia ontology and store in server or with the light weight implementation by performing an endpoint sparql query to retrieve the list of terms.

E. Document Cleaning

The document cleaning function perform two operations. First is to remove the html tag web page. Although there are many html parser libraries such as jsoup[4], HTML parser[5], and others available, however, in this platform we are writing our own parser to increasing flexibility of the configuration and advance testing. The second operation perform by the document cleaning function is to remove unwanted symbol from the web page text. The clean text is very crucial to the document preprocessing process in the next function.

F. Document Statistic

The document statistic function will perform the tf.idf calculation. In order to perform the td.idf, the documents need to go through a set of document preprocessing process such as lexical analysis, elimination of stop words, and stemming. The documents content will first go through the stop words elimination based on a list of stop words for both Malay and English language. Then, the function will perform the stemming operation to get the root word of the content. For English documents we are using the Porter Stemming Algorithm [41]. As for Malay documents, we are using the stemmer that developed in [42]. The result of the tf.idf calculation will be store in a data document for further analysis and reviewed.

G. Document Ranking

The document ranking function is to rank the document result based on the query back to the user. This is to relate the relevancy of the retrieve document to the user query. The ranking task is performed by using a ranking model f (q, d) to sort the documents, where q denotes a query and d denotes a document. Ranking has been widely adopted in IR, data mining, and natural language processing [43]. BM25 and Language Model for IR (LMIR) use a conditional probability model to calculate the relevancy ranking of the document to the ranking [44]. User feedback may be incorporated in future to increase the accuracy of the ranking operation. Some other ranking methods that can be used includes ranking based on ontology [45], ranking based on vector space model [46], ranking based on mean variance analysis [47] and other probabilistic ranking algorithms. These document ranking methods can be adopted in this framework and a comparison analysis can be performed to analyze the performance of these methods.

IV. Conclusion

This paper presented a robust framework for information retrieval for both generic and specific search. However, there is still room for improving this framework. Our team is working to incorporate more function so that the framework becomes more robust. Besides that, future enhancement such Malay NER and POS will be incorporate to increase the accuracy of the information extraction function. Another of our team member is working on the link analysis to increase the retrieval result and to dynamically determine the depth of searching/crawling. According to Wikibon blog, there are 2.7 Zeta bytes of data exist in the digital universe, 571 new websites are created every minute of the day, 100 terabytes of data uploaded daily to Facebook, all these big data are various form either in structured, semi-structured or unstructured [5]. The amount data information available is definitely more that human can process manually, even want to process manually it would take ages to complete. Thus, information retrieval tools such as the framework we proposed provide a tool let to focus on the relevant information and relax user from the time consuming task in gather information and processing unwanted information.

[4]http://jsoup.org/

[5]http://htmlparser.sourceforge.net/

References

[1] B. Y. Ricardo and R. N. Berthier, *Modern Information Retrieval: The Concepts and Technology behind Search*, ACM Press Books, Addison-Wesley Professional, 2nd edition, 2011.

[2] M. Sanderson, and W.B. Croft, "The history of information retrieval research," *Proceedings of the IEEE Special Centennial Issue*, vol. 100, pp. 1444-1451, 2012.

[3] C. N. Mooers, "The theory of digital handling of non-numerical information and its implications to machine economics," in *Association for Computing Machinery Conference*, Rutger University, 1950.

[4] H. F. Mitchell, "The use of the univ ac fac-tronic system in the library reference field," *American Documentation*, vol. 4, no. 1, pp. 16-17, 1953.

[5] (2012). Wikibon Blog: "A Comprehensive List of Big Data Statistics". [Online]. Available: http://wikibon.org/blog/big-data-statistics

[6] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107-117, 1998.

[7] T. Adi, O. K. Ewell, and P. Adi, *High selectivity and accuracy with readware's automated system of knowledge organisation*, 1999.

[8] N. Guarino, "Formal Ontology and Information Systems," in *Proc. the 1st International Conference on Formal Ontology in Information Systems*, Trento, Italy, IOS Press, Amsterdam, pp. 3-15, 1998.

[9] F. Christos and W.O. Douglas, "A Survey of Information Retrieval and Filtering Methods," Technical Report, University of Maryland at College Park, College Park, MD, USA, 1995.

[10] D. M. Sunday, "A very fast substring search algorithm," *Commun. ACM* Vol. 33, no. 8, pp. 132-142, 1990.

[11] L. DikLun and L. Chun-Wu, "Partitioned signature files: design issues and performance evaluation," *ACM Transaction Information System*, vol. 7, no. 2, pp.158-180, 1989

[12] Z. Justin, M. Alistair, and S. D. Ron, "An Efficient Indexing Technique for Full Text Databases," in *Proc. the 18th International Conference on Very Large Data Bases (VLDB '92)*, Li-Yan Yuan Ed. pp.352-362, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1992

[13] G. Salton and A. Wong, "Generation and search of clustered files," *ACM Trans. Database Syst.* vol. 3, no. 4, pp. 321-346, 1978.

[14] W. B. Croft, "A model of cluster searching based on classification," *Information Systems*, vol. 5, issue 3, pp. 189-195, 1980.

[15] A. Ram, "Interest-Based Information Filtering and Extraction in Natural Language Understanding Systems," in *Proc. the Bellcore Workshop on high-performance information filtering*, Morristown, NJ, 1991

[16] P. W. Folt, "Using latent semantic indexing for information filtering.," *SIGOIS Bull*, 11, 2-3, pp. 40-47, 1990

[17] K. L. Kwok, "A neural network for probabilistic information retrieval," in *Proc. the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '89)*, N. J. Belkin, C. J. van Rijsbergen, Eds. ACM, New York, NY, USA, pp. 21-30, 1989

[18] C.-P. Xia, X.-R. Cheng, Z. Lu, and K. Li, "Ontology-based semantic information retrieval," in *Proc. World Automation Congress (WAC)*, 2010, pp. 183-187.

[19] D. Stephen, E. Nadav, G. David *et al.*, "SemTag and seeker: bootstrapping the semantic web via automated semantic annotation," in *Proc. the 12th International Conference on World Wide Web (WWW '03)*. ACM, New York, NY, USA, 2003, pp. 178-186.

[20] P. Borislav, K. Atanas, O. Damyan, M. Dimitar, and K. Angel, "KIM – a semantic platform for information extraction and retrieval," *Nat. Lang. Eng.*, vol. 10, no. 3-4, pp. 375-392, 2004.

[21] L. Fang and H. Xuanjing, "An intelligent platform for information retrieval," in *Proc. the 2005 Joint Chinese-German conference on*

Cognitive systems, R.-Q. Lu, H. Siekmann, and C. Ullrich, Eds. Springer-Verlag, Berlin, Heidelberg, 2005, pp. 45-57.
[22] H. Xiangji, W. Miao, A. Aijun, and H. Yan-Rui, "A platform for Okapi-based contextual information retrieval," in *Proc. the 29th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*, ACM, New York, NY, USA, pp. 728-728, 2006.
[23] C. T. Sergio, M. F. L Juan, F. H. Juan, P. V. Ramiro, and C. R. Julio, "A proposal for an experimental platform on Collaborative Information Retrieval," in *Proc. the 2009 International Symposium on Collaborative Technologies and Systems (CTS'09)*, IEEE Computer Society, Washington, DC, USA, 2009, pp. 485-493.
[24] M. Jayaratne, I. Haththotuwa, C. D. Arachchi, S. Perera, D. Fernando, and S. Weerakoon, "iSeS: intelligent semantic search framework," *EATIS(2012)*, pp. 215-222, 2012.
[25] Ph. Mylonas, D. Vallet, P. Castells, M. Fernndez, and Y. Avrithis, "Personalized information retrieval based on context and ontological knowledge," *Knowledge Engineering Rev.*, vol. 23, no. 1, pp. 73-100, 2008.
[26] A. Bozzon and P. Fraternali, "Multimedia and multimodal information retrieval," in *Proc. SeCO Workshop*, S. Ceri and M. Brambilla, Eds. vol. 5950 of Lecture Notes in Computer Science, Springer, 2009, pp. 135–155.
[27] G. Salton and J. Michael, *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY, USA, 1986.
[28] G. Russell and G. Lapalme. (January 1, 2005). Automatic Identification of Language and Encoding. *Internet Citation*. [Online]. pp. 21. Available: http://rali.iro.umontreal.ca/Publications/files/AILERussellLapalmePlamondon.pdf
[29] G. Thomas and L. Nedim, "A comparison of language identification approaches on short, query-style texts," in *Proc. the 32nd European conference on Advances in Information Retrieval (ECIR'2010)*, G. Cathal, H. Yulan, K. Gabriella, K. Udo, L. Suzanne, Eds. Springer-Verlag, Berlin, Heidelberg, 2010, pp. 611-614.
[30] N. H. Rais, M. T. Abdullah, and R. A. Kadir, "Bilingual dictionary approach for malay-english cross-language information retrieval," *Journal Communication Computing*, vol. 8, pp. 354-360, 2011.
[31] C. H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems, knowledge and data engineering," *IEEE Transactions*, vol. 18, no. 10, pp. 1411-1428, 2006.
[32] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A framework and graphical development environment for robust NLP tools and applications," in *Proc. the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, 2002.
[33] Calais: Connect Everything. [Online]. Available: http://www.opencalais.com/
[34] M. Henzinger, "Link analysis in web information retrieval," *IEEE Data Engineering Bulletin*, vol. 23, no. 3, pp. 3–8, 2000.
[35] P.R. Kumar, and A. Singh, "Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval," *American Journal of Applied Sciences*, vol. 7, no. 6, pp. 840-845, 2010.
[36] J. Palme, "Information Filtering," in *Proc. ITS'98 Conference*, 1998.
[37] H. Uri, S. Bracha, and S. Peretz, "Information Filtering: Overview of Issues, Research and Systems," *User Modeling and User-Adapted Interaction*, vol. 11, no. 3, pp. 203-259, 2001.
[38] A. Baron, P. Rayson, and D. Archer, "Word frequency and key word statistics in corpus linguistics," *Anglistik: International Journal of English Studies*, vol. 20, no. 1, pp. 41–67, 2009.
[39] V.K. Boo, P. Anthony, "Meta search engine powered by DBpedia," in *Proc. International Conference Semantic Technology and Information Retrieval (STAIR)*, 2011, pp. 89-93.
[40] E. Vorhees, "Query expansion using lexical semantic relations," in *Proc. the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1994)*, Dublin, Ireland,1994, pp. 61–67.
[41] M. F. Porter, "An algorithm for suffix stripping," in *Readings in Information Retrieval*, K. S. Jones, and P. Willett, Eds., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 313-316.
[42] L.C. Leong, B. Surayaini, and R. Alfred, "Enhancing Malay Stemming Algorithm with Background Knowledge," in *Proc. 12th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2012)*, *Lecture Notes in Computer Science* 7458, Springer-Verlag Berlin Heidelberg, 2012, pp. 753-758.
[43] L. Hang, "A Short Introduction to Learning to Rank," *IEICE Transactions on Information and Systems*, E94-D(10), 2011
[44] R. Stephen and Z. Hugo, "The probabilistic relevance framework: bm25 and beyond," *Foundation Trends Information Retrieval*, vol. 3, no. 4, pp. 333-389, 2009.
[45] M. Shamsfard, A. Nematzadeh, and S. Motiee, "ORank: an ontology based system for ranking documents," *International Journal of Computer Science*, vol. 1, no. 3, pp. 225-231, 2006
[46] L. L. Dik, C. Huei, and S. Kent, "Document Ranking and the Vector-Space Model," *IEEE Software*, vol. 14, no. 2, pp. 67-75, 1997.
[47] W. Jun, "Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval.," in *Proc. the 31th European Conference on IR Research on Advances in Information Retrieval* (ECIR '09), B. Mohand, B. Catherine, M. Josiane, and S. Chantal, Eds. Springer-Verlag, Berlin, Heidelberg, 2009, pp. 4-16.

Rayner Alfred was born in Kota Kinabalu, Sabah. He completed a PhD in 2008 looking at intelligent techniques to model and optimize the complex, dynamic and distributed processes of knowledge discovery for structured and unstructured data. He holds a PhD degree in computer science from York University (United Kingdom), a master degree in computer science from Western Michigan University, Kalamazoo (USA) and a Computer Science degree from Polytechnic University of Brooklyn, New York (USA).

Dr. Rayner leads and defines projects around knowledge discovery and information retrieval at Universiti Malaysia Sabah. One focus of Dr. Rayner's work is to build smarter mechanism that enables knowledge discovery in relational databases. His work addresses the challenges related to big data problem: How can we create and apply smarter collaborative knowledge discovery technologies that cope with the big data problem.

Dr. Rayner has authored and co-authored more than 75 journals/book chapters and conference papers, editorials, and served on the program and organizing committees of numerous national and international conferences and workshops. He is a member of the Institute of Electrical and Electronic Engineers (IEEE) and Association for Computing Machinery (ACM) societies.

Gan Kim Soon is currently pursuing his PhD in computer science with the Center of Excellent in Semantic Agents under School of Engineering and Information Technology, in Universiti of Malaysia Sabah, Sabah, Malaysia. The author's research interests include agent and multi-agent, semantic web, neural networks, information retrieval, and evolution computation.

Chin Kim On received his PhD in artificial intelligence with the Universiti of Malaysia Sabah, Sabah, Malaysia. The author's research interests included gaming AI, evolutionary computing, evolutionary robotics, neural networks, image processing, semantics based visual information retrieval, agent technologies, evolutionary data mining and biometric security system with mainly focused on fingerprint and voice recognition.

He is currently working as a senior lecturer at the Universiti Malaysia Sabah in the School of Engineering and Information Technology, Sabah, Malaysia. He has authored and co-authored more than 60 articles in the forms of journals, book chapters and conference proceedings. He is a member of IEEE and IAENG societies.

Patricia Anthony received her PhD in computer science from the University of Southampton in 2003. She is currently working as a senior lecturer at the Department of Applied Computing, Lincoln University, New Zealand. Her research interest is in semantic agents and multi-agent systems and how these agents can interact with each other within an open domain to solve problems. She is also interested in investigating how agents can communicate with each other at the semantic level using semantic technology. To date, she has published more than 80 articles in the forms of journals, book chapters and conference proceedings. She is a member of IEEE, ACM and IACSIT.