Information Retrieval for Malay Text: A Decade Review of Research (2008-2019)

Syarifah Fatem Na'imah Binti Syed Kamaruddin
Faculty of Ocean Engineering Technology and Informatics
Universiti Malaysia Terengganu
Kuala Nerus, Terengganu, Malaysia
fatem_alyahya@yahoo.com

Fatihah Mohd
Faculty of Entrepreneourship and Business
Universiti Malaysia Kelantan
Pengkalan Chepa, Kota Bharu, Kelantan, Malaysia
fatihah.m@umk.edu.my

Mohd Pouzi Hamzah
Faculty of Ocean Engineering Technology and Informatics
Universiti Malaysia Terengganu
Kuala Nerus, Terengganu, Malaysia
mph@umt.edu.my

Fadilah Harun
Faculty of Ocean Engineering Technology and Informatics
Universiti Malaysia Terengganu
Kuala Nerus, Terengganu, Malaysia
fadilahharun@unisza.edu.my

Noor Raihani Zainol
Faculty of Entrepreneourship and Business
Universiti Malaysia Kelantan
Pengkalan Chepa, Kota Bharu, Kelantan, Malaysia
raihani@umk.edu.my

Nurul Izyan Mat Daud
Faculty of Entrepreneourship and Business
Universiti Malaysia Kelantan
Pengkalan Chepa, Kota Bharu, Kelantan, Malaysia
izyan.md@umk.edu.my

2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP) | 978-1-6654-1237-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/CAMP51653.2021.9498034

***Abstract*—In this paper, we survey and classify most of the information retrieval (IR) approaches to Malay text in order to assess their benefits and limitations. We also summarized the information retrieval tools and related methods, in which ontology is a widely used tool for all countries' researchers. This research selects Malay language as the primary test collection because there are more issues in Malay languages, particularly those related to deep semantics, including the use of ontology. The traditional Malay retrieval system mostly focused on syntax extraction and keywords only. Mostly this technique will ignore the semantic element and the real meaning of query text and corpus which not fulfil the requirement of the user. Most of the previous study in information retrieval was using English and Arabic language as a test collection. Therefore, advance research is needed and it will be experimented in the future work. The finding of the paper will help other researchers discover the information and research gap regarding the Malay text.**

Keywords— Information retrieval; knowledge, Malay text, natural language processing, ontology

I. INTRODUCTION

One of the ways human obtains information is through the mediation of others who are experts or experienced. Based on their experience and knowledge, their expertise will provide the references or documents related to the subject area requested. When the information seeker examines the documents, they will update their knowledge and may also share with the experts about the documents relevant to the subject area sought. Knowledge is an essential intellectual asset for an organisation, whether public or private. Knowledge also refers to the understanding of a subject area. It includes concepts, facts, and relationships between them as well as mechanisms on how to combine them to solve problems in the subject area [1]. In facts, knowledge is different from data and information. Data are facts about an entity, while information is the relationship between those facts.

Data or documents in an organization are becoming more digital and systematic as information technology advances. It is critical that the document's explicit and implicit knowledge can be processed by a computer using the information retrieval method. Information retrieval is defined as a process of finding documents relevant to a request provided by a user from a collection of documents. It also deals with the representation, storage, organization and indexing of documents. The primary goal is to complete all documents relevant to the user's needs.

Fig. 1 shows the workflow of information retrieval. While, Fig. 2 display how the information work to form new knowledge. The new knowledge could also assemble two or more pieces of existing knowledge or logical connection between them in appropriate ways or techniques to produce hidden knowledge.

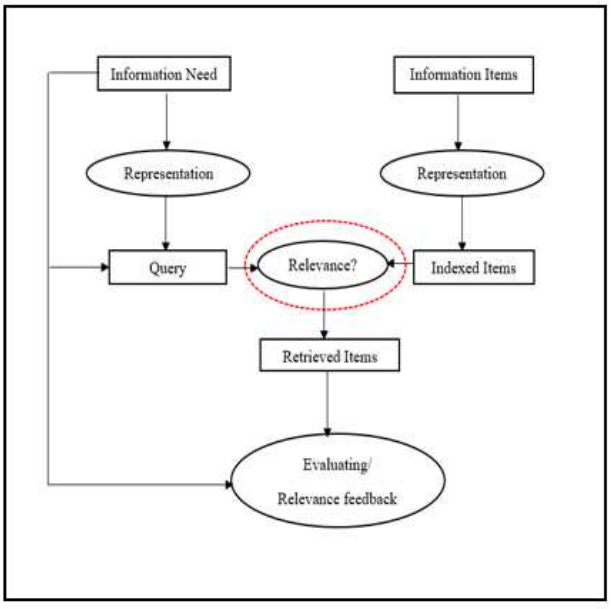

Fig. 1. Workflow of information retrieval

The rest of this paper is organized as follows. Section II describes the information retrieval which is including process, area and evaluation measure of information retrieval

system. Section III presents and discuss the finding of information in this study. Section IV is the conclusion to the study.

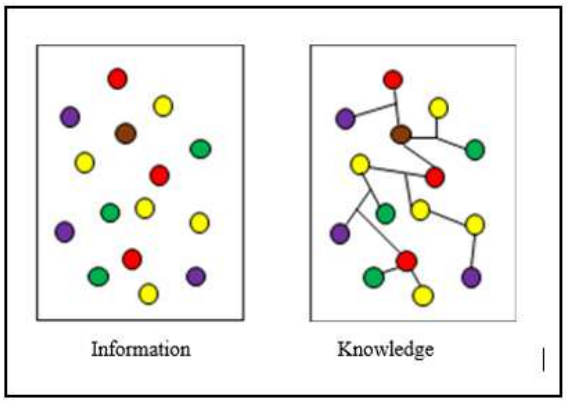

Fig. 2. Example of information and knowledge

II. INFORMATIONA RETRIEVAL SYSTEM

Recently, increasing volume of documents will make the task complicated for us to extract the relevant information or knowledge for the user. Based on the studies conducted by the Dewan Bahasa dan Pustaka (DBP), Malaysia, the final count of the texts collected in the DBP database has exceeded 120 million words of writing text which includes the old Malay texts (from the *Hikayat* and *Kitab*) and the modern texts taken especially from the source of books, newspapers, and magazines. Obstacles and challenges by users to obtain relevant and useful information are increasing with increased data. So, information retrieval system helps the user to retrieve a relevant document and rank them.

This section discusses about the information retrieval system to give more information and find a research gap in the information retrieval field. Information retrieval (IR) is defined as a process of finding documents (information) relevant to a request (generally a query) provided by a user from a collection of documents [2]. It also deals with the representation, storage, organisation, and indexing of documents. The main goal is to achieve all the documents relevant to the needs of the user [3]. The scope of information retrieval also covers other areas such as information management systems, decision support systems, database management systems, and natural language processing.

A. Process in Information Retrieval

The basic information retrieval process consists of three main modules, namely the textual representation of documents, representation of user needs, and comparison between the two representations. Guided by Rijsbergen, the information retrieval system architecture in Fig. 3 can explain this process [4]. The text in a document stored in a text database will be processed by text operations and converted to a logical form. It will then be followed by an indexing process to form a text representation. The user will specify his requirements through the user interface, which will then be broken down and converted to logical form by text operations. The logical structure of this user need will be processed by the query operation, which will generate the actual query or even a representation of the user's needs. This query will then be processed to retrieve the relevant document based on the previously constructed text index. The level of relevance will determine these documents through a ranking process before being sent to the user. Users will examine documents organised according to their degree of significance and find information related to their needs. Chances are, at this point, the user will find a list of documents that have a low degree of relevance but felt significant, then the user will start a feedback cycle. The system will use user-selected documents to reformulate the query in hopes of getting a better representation of user needs.

From this information retrieval process, it found that the information retrieval system can be expand by improving the quality of textual representation documents. The keyword index-based document representations alone cannot provide an accurate picture of the subjects discussed in the paper. One of the problems that made information search difficult is users are searching for information using descriptors differ from the descriptors used in the document. Thus, the results provided by information retrieval systems based on keyword indexes alone are not able to offer effective outcomes. This is due to the techniques used, which reach a lot of irrelevant documents, while many relevant documents are left out [5].

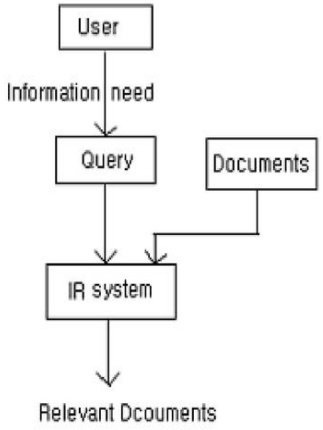

Fig. 3. Information retrieval process

B. Area of Information Retrieval

This section discusses the related research areas of information retrieval such as ontology, semantic and natural language processing. One of the techniques to extract the knowledge in information retrieval is using an ontology approach [6].

Document representation through ontology makes an information retrieval system intelligent. This is because ontologies can store meaning to words and are able to make inferences through semantic relationships between phrases by enabling inference through inheritance [7]. This makes ontology, one of the most popular and powerful tools in knowledge representation [8]. Ontologies can also convert knowledge in unstructured texts into structured forms. This structured knowledge can be understood and processed by computers to be applied in a variety of fields. There is a lot of research in ontology development to enable knowledge to be shared and reused [9]. Therefore, this is an advantage in information retrieval systems to find more accurate document representation methods and produce a much effective information retrieval system.

Another approach that is related to information retrieval is Natural Language Processing (NLP). The task in NLP

must be considered in order to obtain more relevant documents and effectiveness in the information retrieval process. NLP is a process to analyse texts and derive the meaning and understanding of a human language [10]. There are several tasks usually use in NLP such as syntax analysis including tokenization, part-of-speech (POS), lemmatization, stemming, morphology, word sense disambiguation and others [11]. Named Entity Recognition (NER) is the fundamental task in NLP that usually used in information retrieval system [12]. Mahmood used a NER approach to extract named entities and classification by an ensemble-based algorithm. Another researcher also uses the NLP approach in Italian language which use a rule-based to extract relevant documents for clinical records based on their attributes and relations expressions [13]. Another researcher used hybrid method to extract NER, which is a combination machine learning and ruled-based to obtain more accurate recognition to extract names of people, organizations, and locations [14]. However, this method has problem to identify the correct entity that is caused by ambiguity to detect entity boundary. In addition, this method fails to detect location and organization if the adjacent words are not in the list.

Besides information retrieval, the previous researcher also uses knowledge representation approach to extract the knowledge such as semantic nets, system architecture, frames, rules and ontology [15-17]. In certain system, it may apply the rules to the knowledge based in order to extract the knowledge and solve the problems [18-19].

C. Evaluation Measure

This section discusses an evaluation measures that commonly used by previous researchers in information retrieval. Precision, recall and F-measure are the way to calculate the relevant documents. Precision is the fraction of retrieved documents that are relevant and recall is the fraction of relevant documents that are retrieved. The formula shows the precision, recall and F-measure:

$$\text{Precision} = \frac{\text{relevant items retrieved}}{\text{retrieved items}}$$

$$\text{Recall} = \frac{\text{relevant items retrieved}}{\text{relevant items}}$$

$$\text{F-measure} = \frac{2.\ \text{precision.\ recall}}{\text{precision recall}}$$

Table I shows the example of evaluation measures were done from the previous researchers. The increasing percentage of precision and recall state the improvement and effectiveness of retrieval approach.

TABLE I. EVALUATION MEASURES

Precision	Recall	Corpus	Number of queries	References
0.71	0.14	82 docs	49	[1]
0.55	0.14	100 docs	30	[2]
0.75	0.89	Al-Quran	13	[3]
0.90	0.82	32606 tokens	-	[4]
0.65	0.75	50000 data	-	[5]

D. Test Collection

Language is an important element in IR research. This study focuses on extracting from a Malay text. Malay language has four main classes which are noun, verb, adjective and adverb. A linguistic study of Malay words and grammatical structures will be required before extracting the most appropriate structures for common Malay sentence forms. There are several issues in Malay Language because completed Malay dictionary still not exist. In the Malay language experiment, they only depend from the lookup list to identify Malay noun. So, the result is not accurate if the word is not in the list.

There are various test collections that have been developed in different languages; Chinese language used by Di [22] in research of named entity recognition, a collection of Persian language test developed by Ahmadi [14] and the collection of Malay language test used by Sazali [10] and Chekima [23]. The example of the Malay language used by Sazali is a classical Malay text as a collection document.

III. FINDINGS

In order to describe the increasing of the information retrieval domain, this paper aims to review the previous research done in the area of information retrieval and related tasks such as knowledge representation, natural language processing, semantic and information extraction by presenting an information retrieval research practices published in the Scopus for one decade from 2008 to 2019. For this review, the Scopus database is used as a search engine. In this paper, several steps were used to retrieve publications by using “information retrieval” as the main keyword. The keyword is very important in research because it will affect the information on research trend. The research trends could be evaluated by analysis on the frequency of words in title, words in abstract and author keywords in different periods. We trace over 24,732 articles using main keyword and 2,989 related articles after the filter in the area of semantic. To review the research for the past 10 years in this area, we analyse by grouping the research publications into seven (7) main continents, include Asia, North America, South America, Europe, Middle East, Australia and Africa. Table II presents the result of research publication according continents, where Europe is a highest publication with 35.90% followed by Asia with 28.81%.

TABLE II. RESEARCH PUBLICATION BASED ON CONTINENTS

Geographical Areas	Publications (%)
North America	19.14
Asia	28.81
Europe	35.90
Middle East	5.05
Australia	2.01
South America	2.98
Africa	1.74
Undefined	4.39

We will discuss about the productivity among the country. Fig. 4 shows the number of articles up to 15 countries that published from 2008 until 2019 in the information retrieval field for the computer science area. The frequency of publication is important to show the significance and interest in the research area. The data demonstrate that, China and United Stated are the major contributor to the information retrieval field, which are

20.46% and 18.23% respectively. Malaysia is the 17th contributor with a total of 977 articles or 1.56%. This finding shows that Malaysia is lagging behind other countries that produce many articles throughout the year.

The authors in the articles also have a role as a reference to the researcher. We can refer the main author as an expertise in the research area. Table III illustrates the active author in the information retrieval field. The majority of the authors come from China and United States. There are three countries from Asia that able to contribute many publications in this field which are China, India and Japan.

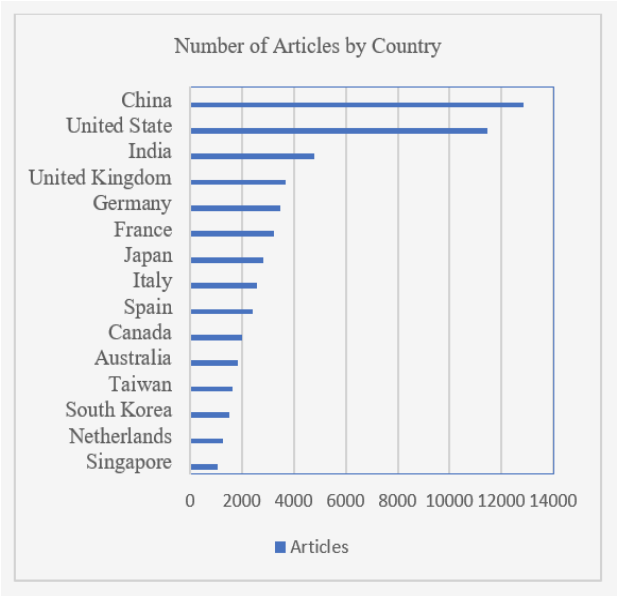

Fig. 4. Number of articles by country

TABLE III. LIST OF AUTHORS

Authors	Publications	Country
De Rijke, Maarten	135	Netherlands
Jones, Gareth J.F.	120	Ireland
Ounis, Iadh	101	United Kingdom
Azzopardi, Leif A.	100	United Kingdom
Jose, Joemon M.	94	United Kingdom
Kamps, Jaap	83	Netherlands
Boughanem, Mohand	81	France
White, Ryen W.	78	United States
Zuccon, Guido	78	Australia
Crestani, Fabio	77	Switzerland
Müller, Henning	77	Switzerland
Tian, Qi	74	United States
Macdonald, Craig	72	United Kingdom
Song, Dawei	69	China
Croft, W. Bruce	68	United States

In this section, we also summarized the Information Retrieval tools, prevalent and related methods. We provided a list and overview of IR methods. In addition, the methodologies prevalently adopted in IR were demonstrated. One of the techniques used in IR is an ontology. In computer science, an ontology is a frequent approach to retrieve a new knowledge of the test collection, such as Word Net, RDF and OWL.

Recently, most of the ontology technique and query languages have been implemented and still on going. The ontology language is important and widely use in ontology-based system. To apply ontology concepts in query formulation, the evaluation of the expressive power, tools and reasoning support is needed to evaluate and choose the best ontology language in the process. Ontology language is important for the user to decide when implement the ontology-based system. Most of the ontology language based on the Extensible Markup Language (XML) which allow them to be machine interpretable. There are several examples such as the Resource Description Framework (RDF) and RDF Schema,the DARPA Agent Markup Language and the Ontology Inference Layer (DAML + OIL) and the Ontology Web language (OWL) and OWL2. OWL is the better language to support for semantic expression compared to RDF. Table IV shows the research that already uses ontology approaches [24].

TABLE IV. ONTOLOGY APPROACHES

Approach	Ontology Technology	Year
Conceptual Graph	WordNet	2002
Ontology Driven Semantic Search	RDF	2004
Vector Space Model for Ontology Based IR	RDF, RDQL Query	2007
Comparison between Classical IR and Ontology Based	RDF, OWL and DAML	2009
Ontology Based Ranking of Web Search Engine	WordNet	2012
Ontology Based Semantic Search Engine	RDF	2017
Knowledge Modelling and Information Retrieval	XML	2017

Besides that, Machine Learning is also one of the popular methods in IR especially topic modelling. Machine learning is divided into supervised and unsupervised. Topic modelling is one of the unsupervised machine learning technique because it doesn't use a training data that already tags or classified by human and expert. Topic Modelling also popular in another field such as text mining, knowledge representation and information extraction. Topic modelling can detect a hidden knowledge in a text by use text mining tool. Its able scanning the documents for abstract the topics, capable of detecting a phrase patterns and also clustering the words by their group.

In topic modelling, a topic will be calculated as a probability distribution compared use a single vocabulary before this. Topic modelling have several techniques which are Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM), Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Non-Negative Matrix Factorization.

In 2015, Gao [25] represent the new model based on relevance and useful information by user for identifying a new pattern that improved the topic model. The earlier of a topic model in 1990 that proposed a Latent Semantic Indexing technique by Deerwester [26]. Then, from The LSI model, it improved the model to probabilistic topic model (PLSA). Blei [27] extend the PLSA model for Latent Dirichlet Allocation (LDA) with the more comprehensive probabilistic model. Most of this model applies in the text analysis for unsupervised topic discover. There are

increasing a number of probabilistic models by integrating several tasks based on LDA.

Lastly, this section also discusses the findings of the related research focus on semantic in information retrieval for Malay language. These findings will review the past studies to stimulate the new knowledge and generate a research gap. Most of the previous study in information retrieval was using English and Arabic language as a test collection.

This research chooses a Malay language because there are more issues in Malay language especially related in semantic. Based on Table V, there are only 68 articles were using Malay language and categorize that according to country where 23 articles from Malaysia. Besides that, semantic area also a main element in this discussion especially for a deep semantic. Currently the techniques to retrieve relevant document for users very limited studies that focus on the deep semantic of Malay language on topic modelling in information retrieval.

TABLE V. RESEARCH PUBLICATION BASED ON COUNTRY THAT USE MALAY LANGUAGE

Country	Publications
Malaysia	23
United States	6
Spain	5
Australia	4
Singapore	4
France	3
Indonesia	3
China	2
India	2
Japan	2
Netherlands	2
United Kingdom	2
Belgium	1
Benin	1
Canada	1
Egypt	1
Iran	1
Italy	1
Jordan	1
Mexico	1
Nepal	1
Pakistan	1

IV. CONCLUSION

The paper has analysed the research in information retrieval field. This field is dominated by Chinese and American countries, with India coming in third where English is its main test collection. Therefore, the English text collection is more advanced than Malay text. There are still have a syntax issue in Malay text study during the analysis process, especially in labelling, segmentation and word disambiguation [22]. As a result, more research of semantic area is required in Malay text, and it will be tested in future work.

ACKNOWLEDGMENT

We would like to appreciate Universiti Malaysia Terengganu and Universiti Malaysia Kelantan for their support in completing this research.

REFERENCES

[1] C. Zins, "Conceptual approaches for defining data, information, and knowledge," Journal of the Association for Information Science and Technology, vol. 58(4), pp. 479-493, 2007.

[2] T. M. T. Sembok, "Knowledge Representation in Information Retrieval," pp. 149-155, 2015.

[3] Z. A. Merrouni, B. Frikh, and B. Ouhbi, "Toward Contextual Information Retrieval: A Review And Trends," Procedia Computer Science, vol. 148, pp. 191-200, 2019.

[4] C. J. V. Rijsbergen, "Information Retrieval," 2nd edn. London, Butterworths, vol. 11(3), pp.237-237, 1979.

[5] H. K. Azad, and A. Deepak, "Query expansion techniques for information retrieval: A survey," Information Processing & Management, vol. 56(5), pp. 1698-1735, 2019.

[6] J. Paralic, and I. Kostial, "Ontology-based information retrieval", Information and Intelligent Systems, pp. 23-28, 2003.

[7] L. Stanchev, "Semantic Document Clustering Using Information from WordNet and DBPedia," Book Semantic Document Clustering Using Information from WordNet and DBPedia, pp. 100-107, 2018.

[8] K. Munir, and M. S. Anjum, "The use of ontologies for effective knowledge modelling and information retrieval", Applied Computing and Informatics, vol. 14(2), pp. 116-126, 2018.

[9] B. Selvalakshmi, and M. Subramaniam, "Intelligent ontology based semantic information retrieval using feature selection and classification," Cluster Computing, vol. 22(5), pp. 12871-12881, 2019.

[10] S.S. Sazali, N. A. Rahman, and Z. A. Bakar, "Information extraction: Evaluating named entity recognition from classical Malay documents," Book Information extraction: Evaluating named entity recognition from classical Malay documents, pp. 48-53, 2017.

[11] T. Lan, and R. Logeswaran, "Challenges And Development In Malay Natural Language Processing," Book Challenges And Development In Malay Natural Language Processing, pp. 61-65, 2020.

[12] A. Mahmood, H. U. Khan, U. R. Zahoor, and W. Khan, "Query based information retrieval and knowledge extraction using Hadith datasets," Book Query based information retrieval and knowledge extraction using Hadith datasets, pp. 1-6, 2018.

[13] C. Diomaiuta, M. Mercorella, M. Ciampi, and G. D. Pietro, "Medical Entity and Relation Extraction from Narrative Clinical Records in Italian Language," Book Medical Entity and Relation Extraction from Narrative Clinical Records in Italian Language, pp. 119-128, 2017.

[14] F. Ahmadi, and H. Moradi, "A hybrid method for Persian Named Entity Recognition," Book A hybrid method for Persian Named Entity Recognition, 2015.

[15] C. M. D. O. Rodrigues, F. L. G. D. Freitas, and R. R. D. Azevedo, "An Ontology for Property Crime Based on Events from UFO-B Foundational Ontology," Book An Ontology for Property Crime Based on Events from UFO-B Foundational Ontology, pp. 331-336, 2016.

[16] T. Xu, D. W. Oard, T. Elsayed, and A. Sayeed, "Knowledge representation from information extraction," in Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, Pittsburgh PA, PA, USA, pp. 475-475, 2008.

[17] S. Roychoudhury, V. Kulkarni, and N. Bellarykar, "Mining enterprise models for knowledgeable decision making," in Proceedings of the Fourth International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering, Florence, Italy, pp. 1-6, 2015.

[18] Y. Wang, "Research on the construction of ontology-based criminology knowledge base," Book Research on the construction of ontology-based criminology knowledge base, pp. 123-128, 2010.

[19] Y. Chouni, M. Erritali, Y. Madani, and H. Ezzikouri, "Information retrieval system based semantique and big data," Procedia Computer Science, vol. 151, pp. 1108-1113, 2019.

[20] N. A. Rahman, Z. A. Bakar, and N. S.S. Zulkefli, "Malay document clustering using complete linkage clustering technique with Cosine Coefficient," Book Malay document clustering using complete

linkage clustering technique with Cosine Coefficient, pp. 103-107, 2016.

[21] R. Othman, and F. A. Wahid, "Quranic texts retrieval in Indri," Book Quranic texts retrieval in Indri, p.p.1-4, 2014.

[22] Y. Di, W. Song, H. Wang, and L. Liu, "Research on open domain Named entity recognition based on Chinese query logs," Book Research on open domain Named entity recognition based on Chinese query logs, pp. 40-44, 2017.

[23] K. Chekima, R. Alfred, and K. O. Chin, "Rule-based model for Malay text sentiment analysis," Book Rule-Based Model for Malay Text Sentiment Analysis, pp. 172-185, 2018.

[24] A. S. Ramkumar, and B. Poorna, "Ontology based semantic search: an introduction and a survey of current approaches," Book Ontology Based Semantic Search: An Introduction and a Survey of Current Approaches, pp. 372-376, 2014.

[25] Y. Gao, Y. Xu, and Y. Li, "Pattern-based topics for document modelling in information filtering," IEEE Transactions on Knowledge and Data Engineering, vol. 27(6), pp. 1629-1642, 2015.

[26] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," vol. 41(6), pp. 391-407, 1990.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3(Jan), pp. 993-1022, 2003.