Open Text Ontology Mining to Improve Retrievals of Information

Mohd Pouzi Hamzah[1], Syarifah Fatem Na'imah Syed Kamaruddin[2]
Faculty of Ocean Engineering Technology and Informatics
University of Malaysia Terengganu
Terengganu, Malaysia

***Abstract*—Information retrieval is the main task to extract relevant information from documents. Mostly, the information retrieval system is based on the keyword approach to extract the knowledge of relevant documents. The experiment shows the ontology can improve the result to overcome the weakness of keyword approach. Ontology implementation method is based on phrase formation and semantic relationships between words. This study tested 10 Malay documents using ontology to retrieve information. The results obtained were compared with the result obtained from manual information retrieval done by experts for precision and recall measure. In this study, there are three semantic relationships between words that are capable of expressing knowledge in documents. They are taxonomy relationship, attribute relationship and non-taxonomy relationship. The relationship of ontology can be formed by using taxonomy relationships algorithm, attribute relationships algorithm and non-taxonomy relationships algorithm based on the linguistic rules of the Malay language. The result of precision and recall for this experiment shows that the ontology approach can enhance the performance of information retrieval from the relevant documents.**

Keywords*—*Information retrieval; ontology; Malay text; taxonomy relationship; non-taxonomy relationship

I. INTRODUCTION

An increase in volume of documents will make the task to extract relevant information or knowledge complicated for users. Obstacles and challenges faced by users to obtain relevant and useful information increases with increased data. Information retrieval system helps users to retrieve a relevant document and rank them. Information retrieval is a process that extracts relevant document from an unstructured document that is meaningful to the user. There are four levels of processing in information retrieval. They are string processing, morphological processing, syntactic processing and semantic processing [1].

We need a data in order to process information and to stimulate knowledge. Data is referring to a fact that can be used in calculating, analysing or processing. The data then becomes information once it has been processed into a form that is useful and meaningful to the user. From the information, we can extract and synthesis more knowledge.

There are some basic elements required in information retrieval systems which are document, query and related comparisons between document and query. Van Rijsbergen [2] has presented the process of information retrieval system as illustrated in Fig. 1. The information retrieval system will start from the input section, which is a query and a document. This query is an interaction between a user and a computer.

The previous researcher also use knowledge representation approach to extract knowledge such as semantic nets, systems architecture, frames, rules, and ontology [3]-[6]. Understanding the limitations of keyword-based information retrieval, this study seeks solutions through knowledge based on documents using ontologies approach. It has been proposed that Malay documents develop ontologies so that knowledge representation of the document can be made. There are implicit relationships between words in the form of phrases, as well as semantic relations between words. These include taxonomic relationships, attribute relationships, and non-taxonomic relationships, among others. Natural language processing is used to establish these semantic relations. They are based on the Malay linguistic rules.

In this paper, we will discuss on the effectiveness of ontology approach in information retrieval. Ontology is a model that explains the world or a particular subject field that consists of a set of properties and a set of relationships between them. In 2001, Hendler defined ontology as "a set of knowledge terms, including the vocabulary, semantic interconnections, and simple rules of inference and logic for some particular topic" [7]. Ontology approach in the search process between query and documents provide an interaction between machine and human. Ontology is also able to achieve a relationship between different types of semantic knowledge.

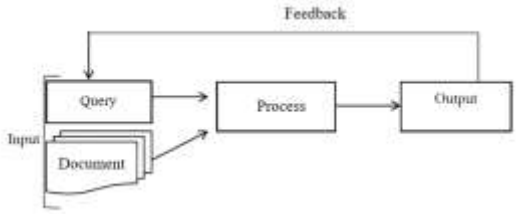

Fig. 1. Information Retrieval Process.

Recently, domain ontologies are also applied in data modelling and information retrieval using semantic-based approaches [8]. The ontology-based approach can improve a semantic gap between the documents and query. The main objective for ontology-based information retrieval is to get more accurate result of the query request by improving the

interface between data and search requests. An ontology-based approach is more efficient in retrieval compared to the tf-idf weighting scheme and latent semantic indexing model [9]. Ontology-based semantic search is one of the search techniques based on the semantic or the meaning of query rather than the syntax of query, and helps to find more relevant information. In general, ontology can be described as in Fig. 2. Fig. 3 shows some of the ontologies in the construction domain. Ontology is very useful because it makes the knowledge in certain subject areas structured and literal. The knowledge can be reused and shared. Due to the richness of semantic information, ontology is widely used in knowledge management systems, artificial intelligence, data mining, knowledge engineering, natural language processing, question and answer systems, information extraction and information retrieval. Chi and Chen [10] used ontology and semantic law to infer knowledge from the characteristics of the news in the document.

The rest of this paper is organized as follows. Section II describes the related work of ontology and Section III represent the Malay test collection. Then, Section VI will cover the methodology of the MyGenOntology. Section V presents and discuss the finding in this study. Section VI is the conclusion to the study.

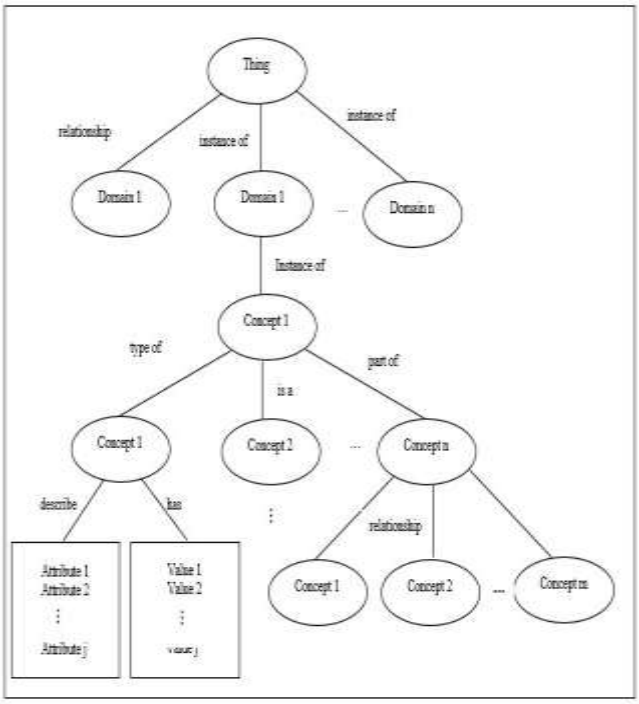

Fig. 2. Ontology.

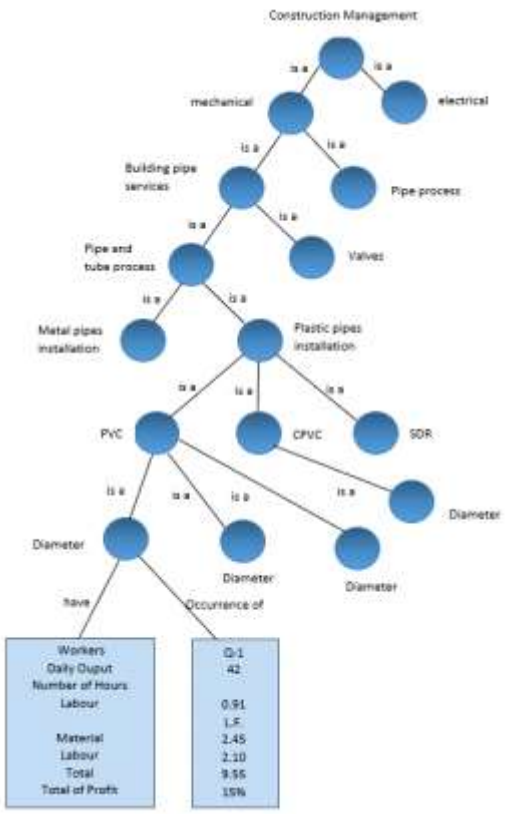

Fig. 3. Ontology of Construction Domain (Source: [11]).

II. Related Work

Most traditional information retrieval systems based on conventional models use only a keyword for document representation. Studies in information retrieval to only keyword based on Boolean models and conventional vector spaces have reached a saturated level with small achievements. Users have difficulty finding the right keywords to get the information they want using Internet search engines such as Google and Yahoo. This difficulty is due to the indexing system used by the search engine. It is based on the keywords found in the document and not based on a concept published from the analysis of the content of the indexed document. However, some search engines have taken the initiative to improve the retrieval process by indexing phrases in documents based on the classification or category of phrase. These methods, however, have their limitations and do not provide encouraging results that fulfil the needs or wants of users, the study has change to semantic aspects that are expected to be able to improve document representation and further enhance relevant document retrieval.

The studies from semantic aspects using thesaurus, phrases, and taxonomies have been extensively reviewed. Although these techniques show improvements, it is still not enough to give high results in information retrieval. Later on, semantic studies are done based on ontology, whereby the representation of documents is based on knowledge organization. The results of this study show that the representation of documents based on ontology can improve the effectiveness and efficiency of information retrieval. However, most of these ontology-based documents representations are made limited to documents in specific domains and not in the general domain.

Representation of a keyword-based cannot express the knowledge of the document. Words with the same meaning (synonyms) are not shown in the representation, while homonyms (words with different meanings) are not distinguished (polysemy). Sánchez [12] states that keyword-based information retrieval models are unable to explain the relationship between phrase and weaknesses in linguistic phenomena such as polysemy and synonyms. Therefore keyword-based information retrieval systems are not able to find relevant documents effectively. Woods [13] stated that two main problems in traditional information retrieval techniques are morphological problems using stemming and semantic problems using query expansion techniques through synonyms. He further explained with the use of "subsumption" technology in which the phrase is arranged in a conceptual taxonomic structure and was tested with specific paragraph retrieval algorithm, it got better results when compared to the results of commercial search engine searches. Using this technology, experiments conducted by Woods [14] on 10 documents obtained a recall value of 38.6% and a precision value of 7.3% compared to the method tf.idf whereby the recall value is 14.8% and the precision value is 2.9%. Yoon [15] who has conducted knowledge-based information retrieval research on specific domains, UMLS ("Unified Medical Language Systems") and SNOMED ("Systematized Nomenclature of Medicine") found that retrieval performance increased by 37% when compared to retrieval performance using traditional vector space.

Research conducted by Yi [16] using ontology-based information retrieval among university students in the United States, got a recall average of 76% when compared to a thesaurus-based information retrieval system of only 43%. Research in the field of information retrieval and knowledge representation based on ontology is accelerating with the advent of semantic web technology. The purpose of semantic web is to make information in the unstructured web meaningful, understandable and can be processed by a computer. The backbone to the semantic web to realize this purpose is ontology. Kumar [17] used an ontology-based semantic indexing approach to show the gap and narrow between text-based websites and semantic websites by using ontology. In 2009, Muhammad [18] proved that stemming algorithm based on stemming order can improve precision and recall. A method based on semantic information can improve the effectiveness of retrieval compared to a method based on keyword.

The methodology of ontology development consists of a set of principles, practices, processes, methods and activities developed to design, construct, evaluate and use ontology [7]. Several ontology development methodologies are reported in the literature. Generally, the development of ontology is made either fully automatic or semi-automatic. Most of the methodologies discussed in the literature use semi-automated methods with a focus on ontological development in specific domains. Uschold and King [19] have developed enterprise ontology for enterprise process modelling. The framework-based methodology of Noy and McGuinness [20] consists of the elements which are identify the domain and scope of ontology, consider the reuse of existing ontology, name the important terms in ontology, define the classes and class hierarchy, define class properties and create class events. The methodology developed by Uschold and Gruninger [21] provides in-depth needs analysis methods and involves four steps which are identify the purpose and scope of ontology, build Ontology, assessment, and documentation. Darlington and Curley [22] used the methodology provided by Noy and McGuinness [20] and Uschold and Gruninger [21] to develop ontologies to support the process of obtaining engineering design requirements. From the Table I below, the precision and recall of ontology is increase from the previous ontology system [23].

TABLE I. PREVIOUS ONTOLOGY SYSTEM

Ontology	Precision	Recall	Accuracy
Google	42.5	100	42.5
Kngine	25	100	25
UsWolfram-Alpha	30	100	30

III. TEST COLLECTION (CORPUS)

Language is an important element in research of information retrieval. There are various test collections that have been developed in different languages; Chinese language used by Di [24] in research of named entity recognition, a collection of Persian language tests developed by Ahmadi [25], and the collection of Malay language tests used by Sazali, Chekima and Sidi [26]-[28]. The example of Malay language used by Sazali is a classical Malay text as a collection document.

This research will focus on the extraction from a Malay text. The texts are collection of Malay texts taken from Berita corpus. The corpus is collected from daily reports and common texts. The corpus is tagged, according to knowledge base. The system will automatically tag proper words according to its class. 10 documents are tagged from a sample of the report at the early stage. They are tagged with a coarse tagset consisting of four main different tags which are noun, verb, adverb and adjective. A linguistic study of Malay words and grammatical structures is required before extracting the most appropriate structures for common forms of Malay sentence.

There are three main systems in language, namely phonological, grammar and semantic systems [29]. Phonological systems are in terms of sound and intonation of language, grammar system is of formation of words and sentences, and semantic system focuses on the meaning contained in language. Grammar system is divided into two main components, which are morphology and syntax. The morphology of the Malay language is a study of two aspects of the language, namely, the process of formation and categorization of words, while the syntax Malay language also is a field of study that examines aspects related to Malay sentences [30]. Since language is a system that has a certain structure, careful analysis of its components needs to be carried out to get meaning from it. To extract meaning, this study focuses on the analysis in the field of morphology and syntax of each sentence in the document.

In Malay language, the sentences have two main parts which are subject and predicate. Subject is the focused matter that is told or described while predicate is the description or story of the focused matter in the subject [31]. The subject section consists of Noun Phrases or other phrases that serve as Noun Phrases. The predicate section may consist of Noun Phrase (FN), Verb Phrase (FK), Adjective Phrase (FA) or Adverb Phrase (FSN). Ramli Md. Salleh [32] stated that in the Malay language, there are four sentence structure bases. They are the first phrase name (subject) + Noun Phrase (predicate), second phrase name (subject) + Verb Phrase (predicate), third phrase name (subject) + phrases Adjectives (predicate) and fourth Noun Phrase (subject) + Adverb Phrase (predicate). In short, construction verses in Malay language consist of four patterns as shown in Table II.

TABLE II. BASIC SENTENCE PATTERN

Pattern	Subject	Predicate
1. FN + FN	Othman	guru sekolah (is a teacher).
2. FN + FK	Syahirah	sedang makan (is eating).
3. FN + FA	Pemuda itu (that young man)	rajin.(is diligent)
4. FN + FSN	Datuk (grandfather)	di kebun.(is in a garden)

IV. METHODOLOGY

Document representation using ontology makes an information retrieval system more effective. This is because ontology can store meaning to words and is able to make inferences through the semantic relationship between words [33]. By enabling inference through hierarchy, ontology is one of the most popular and powerful tools in the representation of knowledge [34]. Ontology can also convert knowledge in unstructured texts into structured forms. This structured knowledge can be understood and processed by computers to be applied in various fields [35]. Therefore, there is a lot of research in the field of ontology development to enable knowledge to be shared and reused [36]. This is no exception in the field of information retrieval to find a more accurate document representation method to produce a more effective information retrieval system.

This section will describe the methods for the development of automatic ontologies for Malay language documents in the general domain. The process will be conducted based on the Malay language system and its development process are combined in a prototype system. This system is called MyGenOntologi. The development of ontology involves the process of extracting ontological components from the text of a general domain without being guided by a pre-defined structure. The extraction system is based on the Malay language system and does not refer to any lexical dictionary. As illustrated in Fig. 4, the ontological components are as follows:

- Concept - An entity or object whether it exists in reality or abstract.
- Attributes - Characteristics or attributes that describe a concept or attributes that describe the identifier of the relationship between concepts.
- Taxonomic Relationships - The hierarchical relationship between concepts.
- Non-taxonomic relationships - Non-hierarchical relationships between concepts.

Fig. 5 show a process of ontology development from texts document. The steps in the ontology development process are as follows:

- Process and analyse documents.
- Extract the phrase from the document.
- Obtain concepts from documents and build taxonomic relationships between concepts.
- Find the concept attributes and attributes to the relationship identifier between the concepts.
- Find non-taxonomic relationships between concepts.

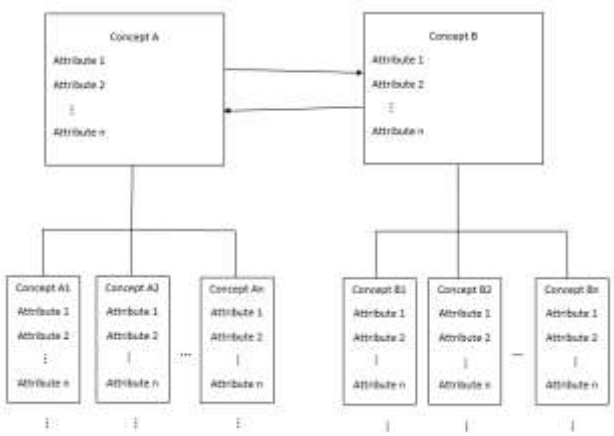

Fig. 4. Ontology Components.

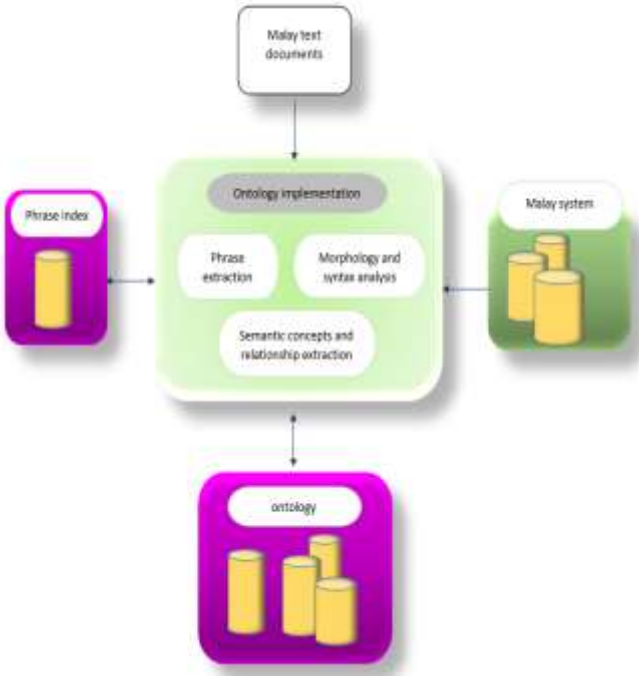

Fig. 5. Process of Ontology Development.

V. Result and Discussion

This section discusses about the issue and the result of the ontology process. There are several errors during the analysis process especially in labelling, segmentation and word disambiguation [24]. These issues are related to the language of corpus. Mostly, the analysis of tagging or labelling will not achieve 100% result because there maybe are spelling errors or the dictionary used does not include certain words. Besides that, the errors may happen because the system cannot detect ambiguous words, or polysemy, and the context of its meaning. Another issue in information retrieval is pertaining to an element of semantic, such as the relationship between words or sentences. These issues can be resolved while improving an ontology or ruled based technique [35].

To identify the effectiveness of ontology quality that has been developed by the MyGenOntology system, the extraction of ontological components from a set of experimental tests consisting of 10 documents from three different online newspapers was carried out. In addition, the manual extraction of ontological components from the same set of experimental tests was also performed by five experts. These manually extracted ontology components are then compared to ontologies developed by the system. Taxonomic relationships show the hypernyms or hyponyms between concepts. The average of recall value is 88% and the precision value achieved is 79%, deeming the experiments conducted as successful. Overall, the recall interval value in percentage is [76,95] and the precision interval value is [62,93]. Comparison of recall value and precision value between documents can be seen more clearly using bar charts as illustrated in Fig. 6.

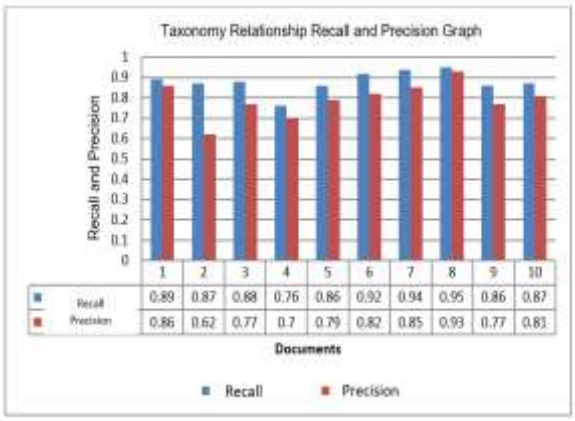

Fig. 6. Taxonomic Relationship Recall and Precision Graph.

The level of precision of the taxonomic relationship is lower than recall. This is because the system forms a taxonomic relationship from the content of the document and the relationship is not included in the list of manually created taxonomic relationships. This problem is due to the weakness during the phrase labelling process and the implementation of taxonomic law. The formation of incorrect compound phrase will affect the form of taxonomic relationships. For example, due to the absence of a comma after the word "beliau" (he) in the sentence “Kata beliau rakyat di Bukit Gantang, Bukit Selambau dan Batang Ai juga...” (He said the people in Bukit Gantang, Bukit Selambau and Batang Ai also ...) from the test collection, the system has produced a phrase of “beliau rakyat” (he people), thus forming a taxonomic relationship between concepts "beliau" (“people”) with the sub-concept "beliau rakyat". In addition, the weakness of the labelling process is due to words that are polysemic. For example, the phrase, “Kesemua pelajar cemerlang terbabit menerima wang tunai RM300, dua buah buku motivasi dan sijil penghargaan” (All the students received RM300 in cash, two motivational books and a certificate of appreciation) resulted in the wrong definition. In the sentence, the word "buah" (a) in the context of the sentence is a collective noun but has been labelled by the system as a noun word causing the system to form a phrase of "buah buku" (a book), thus forming a taxonomic relationship between the concept of "buah" (a) and the sub-concept "buah buku" (a book). Although the recall result is high, there are also taxonomic relationships provided manually that cannot be detected by the system. This is due to several problems. The first problem is the labelling of phrase that are not in the dictionary, such as the taxonomic relationship between the concept of "litar"(circuit) and "litar tertutup"(closed circuit). The word "litar" (circuit) was not found in the Malay dictionary causing the system inability to form the phrase "litar tertutup"(closed circuit). The second problem is the effect of the formation of a compound phrase of two words of the noun in “Chor berkata, ia adalah antara langkah penambahbaikan dari segi pematuhan peraturan dan...”. (Chor said, it was among the improvement measures in terms of regulatory compliance and”.) The system has labelled “segi pematuhan peraturan” (in terms of compliance and regulations) as a compound phrase, causing the taxonomic relationship between the concept of "pematuhan"

("compliance") and the sub-concept of "pematuhan peraturan" (compliance rules) to be undetected manually. Next problem is the grammatical weaknesses. This affects the retrieval of taxonomic relationships. The proper noun "orang kurang upaya" (disabled person) in the sentence "Katanya, 10 peratus lagi adalah untuk pelajar dari Sabah dan Sarawak manakala 10 peratus lagi dikhaskan bagi kes tertentu seperti pelajar orang kurang upaya (OKU)" ("he said another 10 per cent was for students from Sabah and Sarawak while another 10 per cent was reserved for certain cases such as students with disabilities) was written in lower case at the beginning of the sentence. This causes the system failure to identify the concept of "orang kurang upaya oku"(disabled person) and cannot form a taxonomic relationship between "pelajar" ("student") and "pelajar orang kurang upaya oku" ("disabled student). The formation of these incorrect taxonomic relations affects the degree of precision of taxonomic relations.

Another process in the ontology development is the attribute relationship formation. Fig. 9 demonstrates the effectiveness of the formation of the relationship of concepts and attributes. It shows that the effectiveness is moderate with an average recall of 80% and the precision achieved is 54%. Overall, the recall interval value in percentage is [74,90] and the precision interval value is [29,69]. Comparison of values of recall and precision between documents can be seen more clearly using bar charts as in Fig. 7.

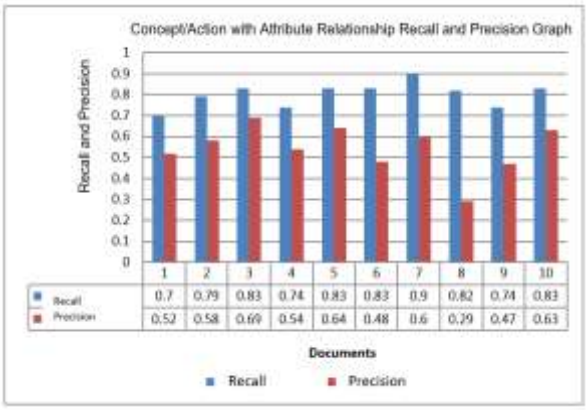

Fig. 7. Attribute Relationship Recall and Precision Graph.

The precision value of the attribute relationship obtained from the tests conducted is low compared to the recall value. This is because the system tries to detect as many phrases as possible that will be attributes to the concept or action in order to increase recall. In general, the weaknesses and problems that affect the precision and recall of taxonomic relationships also affect the precision and recall of the relationship between concepts and actions with attributes. The weaknesses and problems that form incorrect compound phrase will form incorrect attribute relationships, in turn affecting the precision. For example, from the formation of compound phrase "negara tahun"(country year), it will produce an attribute relationship between the concept of "negara"(country). and the concept of "tahun" (year). The problem of labelling a compound phrase from a sequence of more than two nouns as in the sentence "...insiden kematian atau kecederaan orang tahanan berulang" (... incidents of death or injury of detainees are repeated") affects the precision of attribute relationships. This is because the system produces two nouns phrase namely "kecederaan orang" (injuries people) and "tahanan" (prisoner). This makes the attribute relationship between "insiden kematian" ("incident death") and "orang tahanan" (prisoner) and the attribute relationship between "kecederaan" (injuries) and "orang tahanan" (prisoner) undetectable by the system. In the sentence "chor berkata melalui semakan dan kajian yang teliti...", (Chor said through careful review and study ...) the formation of the compound phrase "melalui semakan" (through review") causes the system to not be able to form and detect the attribute relationship between "semakan"(review) and "teliti"(thorough). The normalization problem of the phrase also affects the precision of the test. In the sentence "Bagaimanapun, jelasnya, biasiswa yang ditawarkan kepada pelajar Sabah dan Sarawak adalah khusus untuk bumiputera sahaja" (However, he explained, the scholarships offered to Sabah and Sarawak students are specifically for bumiputera only), the phrase Sarawak is not normalized to "kepada pelajar sarawak" (to Sarawak students) causing the system unable to form and trace the attribute relationship between "ditawarkan" (offered) and "kepada pelajar sarawak" (to Sarawak students).

The next process in the ontology development is the formation of non-taxonomic relationship. Fig. 8 illustrates the effectiveness of detecting non-taxonomic relationships between concepts with an average of recall value of 60% and the precision value achieved is 63%. Overall, the recall interval value in percentage is [38,88] and the precision interval value is [33,100]. Comparison of recall and precision between documents can be seen more clearly using bar charts as in Fig. 8.

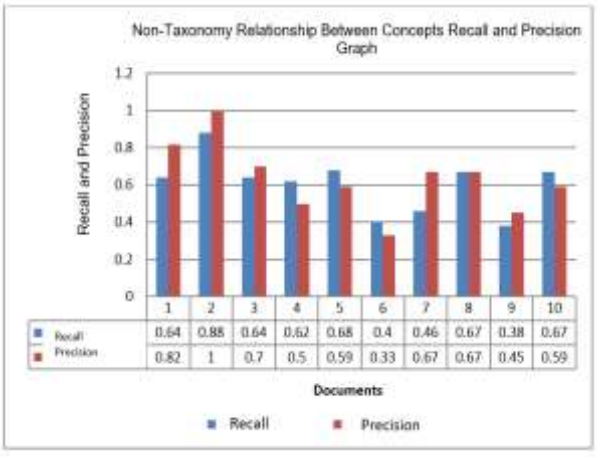

Fig. 8. Non-Taxonomic Relationship Recall and Precision Graph.

Non-taxonomic relationships involve the relationship of three phrases, which are two phrases of nouns and one phrase of verbs, as compared to taxonomic and attribute relationships involving only two phrases. The weaknesses and problems that affect the precision and recall of retrieval to taxonomic and attribute relationships from the process of formation of phrase will affect the precision and recall of non-taxonomic relationship. This is because the taxonomic relationship involves the relationship between the three phrases. The average recall of non-taxonomic relationships is (60%), lower

than the average recall of taxonomic relationships (88%) and attribute relationships (80%). However, the average precision of taxonomic relationship (63%) was higher than the average precision of attribute relationship (54%) but lower than the average prevision of taxonomic relationship (79%). This is because the rules of non-taxonomic relationship search are more limited than the rules of attribute relationship search.

Based on the experiments conducted, it has been found that the cause of inaccuracy of non-taxonomic relationship retrieval is inherited from the factors of weakness and problems that affect the inaccuracy of taxonomic and attribute relationships. The problem of labelling phrase of nouns as in the sentence “Adakah mereka mendapat keperluan nutrisi yang diperlukan?”, (“Do they get the nutritional needs they need?) has caused the system to label "adakah mereka" (are they) as a noun phrase, causing the system inability to form a taxonomic relationship "mereka-mendapat-keperluan nutrisi" (they-get-nutritional needs). Instead, the system formed a taxonomic relationship "adakah mereka-mendapat-keperluan nutrisi" (they-get-the-nutritional needs). In addition to the aforementioned weaknesses and problems, this study also found certain sentence patterns that caused the system to form incorrect non-taxonomic relationships. The words from nouns and verbs are connected with the conjunction word "and". This can be seen in the sentence "Selain itu, pengalaman dan kreativiti guru-guru dalam..." (In addition, the experience and creativity of teachers in ..."). It causes the system to form a non-taxonomic relationship "pengalaman-kreativiti-guru-guru" (experience-creativity-of-teachers).

The results of the experiments found that the effectiveness of the taxonomic relationship between concepts is high with an average recall of 88% and average precision of 79%, the effectiveness of the concept relationship and action between attributes is moderate with an average retrieval of 80% and accuracy of 54%, and the effectiveness of non-taxonomic relationships between concepts is low with an average retrieval of 60% and an accuracy of 63%. Although the effectiveness of the detection of the components of ontologies derived from this experiment is not very high, it is still considered as a success and can be used as a basis for the development of ontology for Malay documents.

VI. Conclusion

In conclusion, there are some important findings have been found from this study. The findings provide a major contribution in the field of information retrieval to Malay documents which are development of detection and generation of taxonomic relations algorithm, attribute relationships algorithm and non-taxonomic relationships algorithm.

Based on the result, it can be concluded that the representation of Malay documents using an ontology-based knowledge repository can help and support information retrieval.

References

[1] T. M. T. Sembok, "Knowledge Representation in Information Retrieval", 2015.

[2] C. J. V. Rijsbergen, Information Retrieval, Butterworth- Heinemann, 1979.

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

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

[5] 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”, Brazilian Conference on Intelligent Systems (BRACIS), 2016.

[6] Y. Chouni, "Information retrieval system based semantique and big data." Procedia Computer Science, vol. 151: pp. 1108-1113, 2019.

[7] D. Gasevic, D. Djuric, and V. Devedzic, “Model Driven Architecture and Ontology Development” Springer, 2006.

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

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

[10] Y. Chi and H. Chen, “Ontology and semantic rules in document dispatching”, The Eletronic Library, vol. 27(4), 694-707, 2008.

[11] C.J. Anumba, R.R.A. Issa, J. Pan, and I. Mutis, “Ontology-based information and knowledge management in construction”, Construction Innovation, vol. 8(3), pp. 218-239, 2008.

[12] M.F. Sánchez, “Semantically enhanced information retrieval: An ontology-based approach”, Tesis Ph.D. Autonoma University De Madrid, 2009.

[13] W. A. Woods, “Searching vs finding: Why systems need knowledge to find what you really want”, ACM Queue, hlm. pp. 2-35, 2004.

[14] W. A. Woods, L. A. Bookman, A. Houston, R. J. Kuhns, P. Martin, and S. Green, “Linguistic knowledge can improve information retrieval”, Proceedings of ANLP-2000, Seattle, WA, hlm. pp. 1-6, 2000.

[15] C. Yoon, “Domain-Specific Knowledge-based information retrieval model using knowledge reduction”, Tesis Ph.D. University of Florida, 2005.

[16] M. Yi, “User performance using an ontology-driven information retrieval (ONTOIR) system”, Tesis Ph.D. Florida State University,2006.

[17] S. Kumar, R. Rana and P. Singh, “Ontology based semantic indexing approach for information retrieval system”, International Journal of Computer Applications, vol. (12), pp. 0975 – 8887, 2012.

[18] M. T. Abdullah, F. Ahmad, R. Mahmod and T. M. T. Sembok, “Rules frequency order stemmer for Malay language”, IJCSNS International Journal of Computer Science and Network Security, vol. 9(2), pp. 433-438, 2009.

[19] M. Uschold and M. King, “Towards a methodology for building ontologies”, IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing. Montreal, 1996.

[20] N. Noy and L. McGuinness, “Ontology development 101: A guide to creating your first ontology”, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

[21] M. Uschold and M. Gruninger, “Ontologies: Principles, Methods and Applications”, Knowledge Engineering Review, vol. 11 (2), 1996.

[22] M. J. Darlington and S. J. Culley, “Investigating ontology development for engineering design support”, Advanced Engineering Informatics, vol. 22(1), pp. 112-134, 2008.

[23] A. Sayed, and A. Al Muqrishi, “IBRI-CASONTO: Ontology-based semantic search engine”, Egyptian Informatics Journal, vol. 18(3), pp. 181-192, 2017.

[24] Y. Di, W. Song, H. Wang and L. Liu, “Research on open domain Named entity recognition based on Chinese query logs”, Proceedings of 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, IMCEC, 2017.

[25] F. Ahmadi and H. Moradi, “A hybrid method for Persian Named Entity Recognition”, Conference on Information and Knowledge Technology, IKT, 2015.

[26] S. S. Sazali, N. A. Rahman and Z. A. Bakar, “Information extraction: Evaluating named entity recognition from classical Malay documents”, International Conference on Information Retrieval and Knowledge Management, CAMP 2016”, 2017.

[27] K. R Chekima, Alfred and K. O. Chin, “Rule-Based Model for Malay Text Sentiment Analysis”, Lecture Notes in Electrical Engineering, vol. 488, pp. 172-185, 2018.

[28] F. Sidi, "MalayIK: An Ontological Approach to Knowledge Transformation in Malay Unstructured Documents." International Journal of Electrical and Computer Engineering (IJECE), vol. 8(1), 2018.

[29] Ahmad Khair Mohd Nor., Pengantar sintaksis bahasa Melayu, Kuala Lumpur: Utusan Publications & Distributions Sdn. Bhd, 2003.

[30] Zulkifley Hamid, Konsep Bahasa, dlm. Zulkifley Hamid, Ramli Md. Salleh, Rahim Aman. (pnyt.). Linguistik Melayu. Bangi: Penerbitan Universiti Kebangsaan Malaysia, 2006.

[31] Abdul Ghalib Yunus, Dr. Ghazali Lateh, Panduan revisi esensi bahasa Malaysia PMR. Longman: Pearson Sdn. Bhd.,2009.

[32] Ramli Md. Salleh, Zulkifley Hamid, Ramli Md. Salleh, Rahim Aman. (pnyt.). Linguistik Melayu. Bangi: Penerbitan Universiti Kebangsaan Malaysia, 2006.

[33] 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.

[34] G.P. Zarri and E. Jennex. Murray, “Knowledge Management: Concepts, Methodologies, Tools, and Applications”, Premier Reference Source, IGI Global, 2008.

[35] 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.

[36] S. B., Rodzman, "Domain specific concept ontologies and text summarization as hierarchical fuzzy logic ranking indicator on malay text corpus."Indonesian Journal of Electrical Engineering and Computer Science, vol. 15(3), 2019.