International Journal of Software Engineering and Its Applications
Vol. 9, No. 5 (2015), pp. 193-204
http://dx.doi.org/10.14257/ijseia.2015.9.5.19

Managing Information by Utilizing WordNet as the Database for Semantic Search Engine

Noryusliza Abdullah, Rosziati Ibrahim

Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, Johor, Malaysia
yusliza@uthm.edu.my, rosziati@uthm.edu.my

Abstract

The growth of data and information has encouraged researchers to overcome information overloaded. One of the convincing methods is through utilization of WordNet, a large lexical database of English. Generally, WordNet is used in Natural Language Processing (NLP) area and increasingly manipulated in the information retrieval field. Some search engines have using the WordNet in acquiring better meaning of user-entered keyword. Its usage might help in addressing the emergence of huge numbers of information on the web. This paper discusses the WordNet as an alternative database for semantic search engine (SSE). The SSE is then used for information retrieval. We tested our semantic search engine and compare it using Google and Yahoo search engines. Results show that our SSE gives better information retrieval when it comes towards personalization using user profiling to produce more relevant results.

Keywords: *WordNet, ontology, semantic search engine, information retrieval, user profiling*

1. Introduction

Information availability is very limited in the past. The privilege in accessing them was allowed only to a certain group of people with high cost engaged. The existence of the World Wide Web (WWW) and search engine has bridging the gap to the information recipients. In line with the maturity of Information Technology, we faced information overloaded. It is estimated that there are over three billions web pages on the internet [1]. Distinguishing between right or wrong and useful or useless information seems to be a prominent task.

Information retrieval is a broad field. It did not begin with the web. Yet, web information retrieval also known as search engine, has become the main resources in this area. The optimization of information retrieval has contributed to the quality enhancement of web search engine [2]. The internet dependency in seeking information has been claimed to growth to 566.4% from 2000 to 2012 [1]. This huge value has proven that internet is a good and accepted place to obtain information.

2. Semantic Search Engine

Primary need to explore the internet is its search engine. Since 1995 until now, search engine is widely exposed to retrieve information on the web. The specific purpose is to search document using keywords. In order to execute the searching, crawler, spider or bot (robot) is used to fetch documents in WWW before extracted by users. This task is conducted in periodical time to obtain changes. Indexer or catalogue will index those documents based on the words. These data are then utilized by search engines to do matching and ranking [3].

While this is proven successful in information retrieval, it is not sufficient to rely on keyword alone in finding the most related web site because substantial number of results

ISSN: 1738-9984 IJSEIA

will cause time incremental to analyze the results. Hence, active researches on search engines have produced several forms of searching techniques to meet users' needs. Table 1 list several searching engines with their searching method and features.

Table 1. Search Engines with their Method and Features

Search Engine	Searching method(s)	Features
Google, Yahoo, Bing	Keyword search	Free
Hakia	Meaning-based search	Semantic search
SenseBot	Text summary for keyword	Use Semantic Web technologies
Powerset	Keyword and phrase semantic meaning	Semantic search

Semantic search is critically needed since search effectiveness in information retrieval is depends on user's characteristics [4]. In that study, search experience and high cognitive skills of the users will give better results. High cognitive skills are defined as perceptual speed, logical reasoning, verbal comprehension and spatial searching. Based on the constraints stated above, it is obvious that only these group of users will benefited by the search engine. If no further action taken, inexperience and low cognitive skills users will be left behind from receiving precise information in acceptable time.

The semantic search engine introduced a concept called ontology. Good search engine that give trusted results required assistance from domain knowledge to answer intelligent queries [5]. Currently, ontology has adopted as one of the reliable domain knowledge. According to Aguilar-Lopez *et al*. [6], extracting semantic of the content by using ontology is the method to address issues of time usage and accurate results. Lexical ontology using WordNet is one of the ontologies branch.

3. Libraries for wordnet

Various methods are applicable in utilizing the WordNet. One of them is using specific libraries or Application Programming Interfaces (APIs). These APIs are capable in accessing WordNet data and manipulate it by applying the available functions. API package consisting of database and libraries and it has to be installed in order to perform the task. Two well-known WordNet APIs in Java programming language are Java WordNet Libraries (JWNL) and Java API for WordNet Searching (JAWS). Different API is used when the application is written in other languages. JWNL [7] is one of the APIs for accessing WordNet data. It can be downloaded for free from the web. Current database is using WordNet data version 3.1. In order to execute web application using JWNL, setting in the web server is established. In this research, Tomcat 7.0 is used as the web server. The web server setting is done as shown in Figure 1.

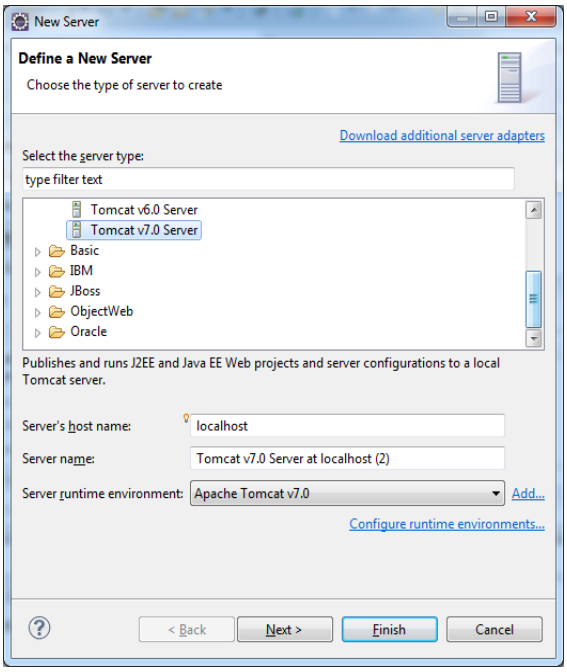

Figure 1. Web Server Setting

To enable the application retrieves the data provided in WordNet, configuration on *file_properties.xml* in config folder is set. It is based on the directory used to store database taken from JWNL package. Figure 2 shows the snippet. In this example, the data is stored in *E:\myPHD\workspace\JWNL\data\wn31\dict*. Package provided by JWNL is completed with all the needed files.

```
            <param name="dictionary_element_factory"
value="net.didion.jwnl.princeton.data.PrincetonWN17FileDictionary
ElementFactory"/>
            <param name="file_manager"
value="net.didion.jwnl.dictionary.file_manager.FileManagerImpl">
                  <param name="file_type"
value="net.didion.jwnl.princeton.file.PrincetonRandomAccessDictio
naryFile"/>
                  <param name="dictionary_path"
value="E:\myPHD\workspace\JWNL\data\wn31\dict\"/>
            </param>
```

Figure 2. WordNet Retrieving Configuration

After the configuration is completed, those data are available through functions in the library. For instance, initializing the WordNet is using *JWNL.initialize* function while *getInstance* function is used in accessing WordNet. Meanwhile the *lookupIndexWord* function is the main lookup procedure to choose word type: Adjactive, Adverb, Noun and Verb.

4. Wordnet as a Database

In Figure 3, WordNet database structure is shown using Toad for MySql 6.3. The database is consisting of 17 interrelated tables. They are composed of nouns, verbs and adjectives.

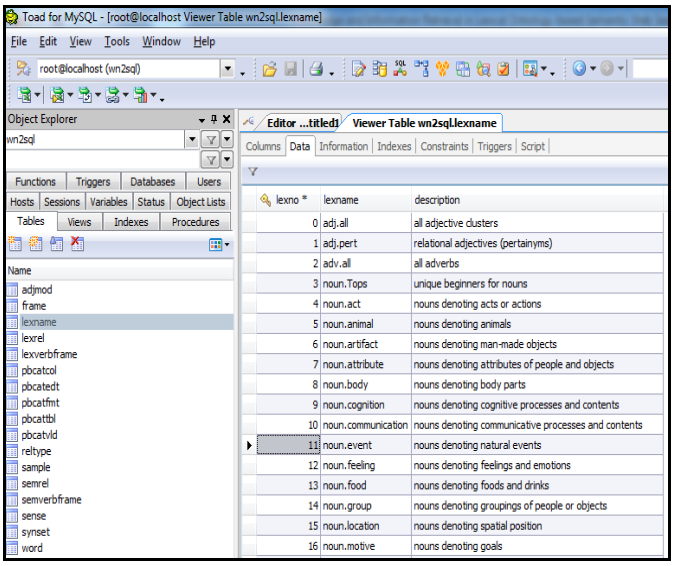

Figure 3. WordNet Database Structure

In keyword searching, usually noun is referred. As an example, Figure 4 indicates sample of WordNet data from four tables. Tables involved are Word, Sense, Synset and Lexname. Table Word posses the word itself (lemma) and word number. It has relationship with Sense table whereby this table list definition for the word. Definition is represented in number. Analyzing the definition from Sense need Synset table. Categories are obtained from Lexname table when a relationship reformed between Synset and Lexname table.

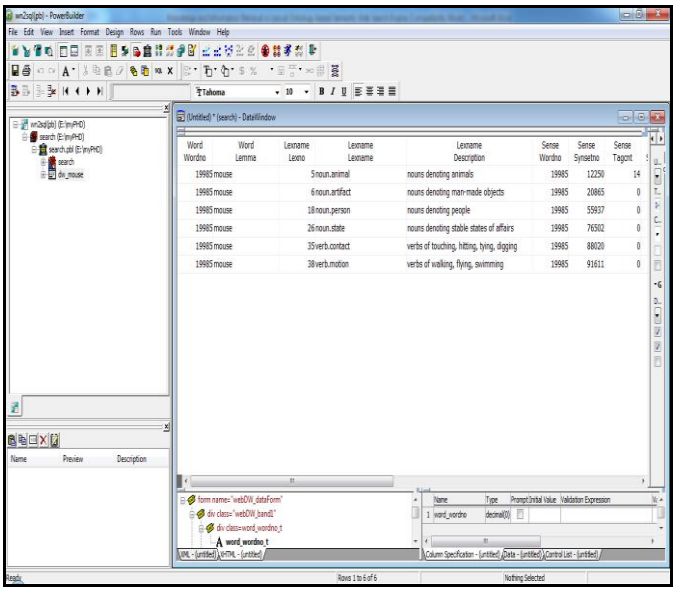

Figure 4. Sample WordNet Data

The relationships of four tables in Wordnet are displayed in Figure 5 in database schema. Those relationships generate categories for certain keyword. Linking between tables to perform relationship will allow in acquiring numerous data from this ontology.

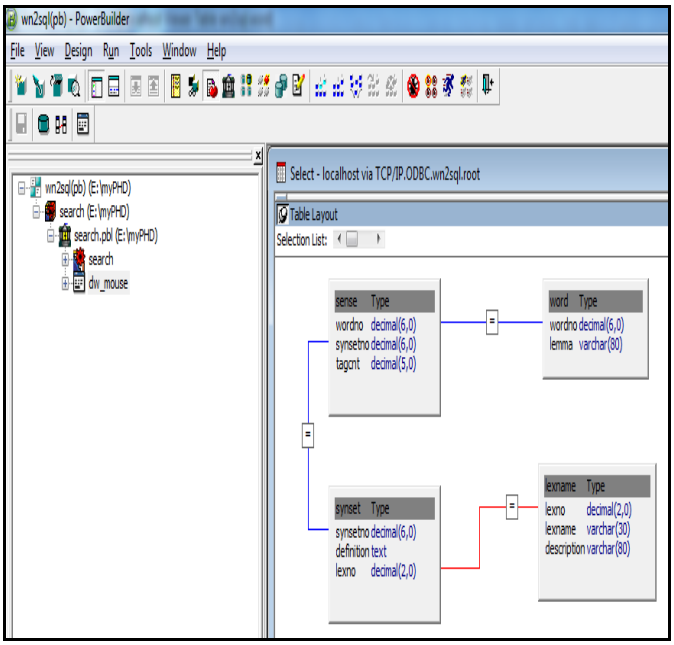

Figure 5. WordNet Database used in the Implementation Phase

In this paper, example of process in finding depth from those data is shown. For the assimilation purpose, 'Java' keyword is used as a sample. WordNet as a lexical ontology is using hypernym-hyponym concept. It acts as a parent-child relationship. Traversing through the hypernyms of word will contribute in depth searching. To facilitate in illustration, every word, sense (and synset) is represented with number. Figure 6 shows the route from word to root for 'Java'.

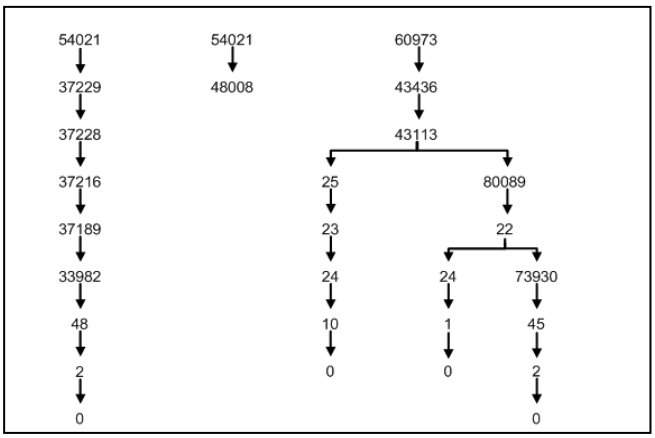

Figure 6. Synset in WordNet with hypernyms

Word 'Java' has three different meaning called sense. Figure 6 shows the three different routes. For simplicity, only Java programming language depth is calculated. SQL is used over WordNet database and recursive technique is utilized in acquiring its depth. Figure 7 is the SQL command.

```
SELECT t1.synsetno1 AS parent,
GROUP_CONCAT(t2.synsetno1  ORDER BY t2.synsetno1 ) AS
node,
COUNT(*)  FROM semrel AS t1
INNER JOIN semrel t2 ON
t1.synsetno2 = t2.synsetno1
where t1.synsetno1 = 43113 and t1.reltypeno in (2,3)
and
t2.reltypeno in (2,3)
GROUP BY parent;
```

Figure 7. SQL Command to Retrieve Depth from Node to Root

Depth can be calculated using this SQL statement for 'Java Programming Language'. The depth value can be used in other applications including similarity measurement between ontologies. The benefit of using this method is the ability in processing complex query. Developers accept it as the strength of SQL. Nevertheless, the technique gives slower results compared to the APIs or libraries.

5. Ontology Representation

Other method to exploit WordNet data is by using it in ontology form. WordNet.owl is available in World Wide Web Consortium site [9]. Web Ontology Language (OWL) is chosen due to the capability in representing and reuse of domain knowledge. This new concept is capable in sharing information structure among people, machines or applications. OWL is selected compared to Resource Description Framework (RDF) and Resource Description Framework Schema (RDFS) because it is better in giving more information about the data. Viewing this file in the form of OWL is shown in Figure 8.

```
<rdf:RDF
xmlns="http://www.ontologyportal.org/WordNet.owl#"
xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl ="http://www.w3.org/2002/07/owl#">
<owl:Ontology rdf:about="WordNet">
<rdfs:comment xml:lang="en">An expression of the Princeton
WordNet ( http://wordnet.princeton.edu ) in OWL.  Use is
subject to the Princeton WordNet license at
http://wordnet.princeton.edu/wordnet/license/</rdfs:comment>
<rdfs:comment xml:lang="en">Produced on date: Mon May 10
00:59:29 PDT 2010</rdfs:comment>
</owl:Ontology>
<owl:SymmetricProperty rdf:ID="antonym">
  <rdfs:label xml:lang="en">antonym</rdfs:label>
  <rdfs:domain rdf:resource="#Synset" />
  <rdfs:range rdf:resource="#Synset" />
</owl:SymmetricProperty>
```

Figure 8. WordNet.owl

Every OWL files can be edited in ontology editors. The editors are Protégé, Top Braid Composer, NeOn Toolkit, SWOOP, Neologism, Vitro, Knoodle and Anzo for Excel. From the above mentioned editors, Protégé [10] is the commonly used editor due to the support of wide variety of plug-in and import formats. Furthermore, it is free open source. However, Wordnet.OWL is failed to open in protégé because of the huge file size.

Data from this ontology is accessible in two ways. First is using Jena Inference. Manipulation using this method needs library called OWL_DL_MEM_RULE_INF. Second method is using SPARQL Protocol and RDF Query Language (SPARQL). It has the ability to do complex query. Although by using inference, results can be obtained in a simplest way, yet, inference engine has not established to provide comprehensive query. On the other hand, SPARQL needs longest time to process query compared to the previous two methods discussed earlier: API and database.

6. Results and Discussion

We tested our SSE against Google and Yahoo as they are widely used search engines. The SSE is developed using user profiling and domain ontology concept by utilizing semantic similarity measurement discussed in [11]. Table 2 shows the characteristics of the search engines.

Table 2. Characteristics of Different Search Engines

Search Engine	Characteristics
Google	PageRank technology
Yahoo	Yahoo Slurp and Bingbot as crawler
SSE	User profiling and domain ontology

Google is an internet-related services and products. Search engine is one of its popular invented products using PageRank technology. Meanwhile Yahoo is the second larger search directory after Google. Previously, Yahoo is using Google's search engine to obtain results before shifting to Yahoo Slurp and the latest crawler is Bingbot. Google and Yahoo Search have been giving ultimate benefits to internet searchers since 1997 and 2001 respectively. However, distinct features in SSE have advantages in terms of ontological concept, categorization and user profiling. It is capable to give categorized and personalized results.

Comparison is done between SSE, Google and Yahoo. Google is giving 221 millions of results when 'Java' keyword is entered as shown in Figure 9 and Figure show 179 millions of results provided by Yahoo search engine.

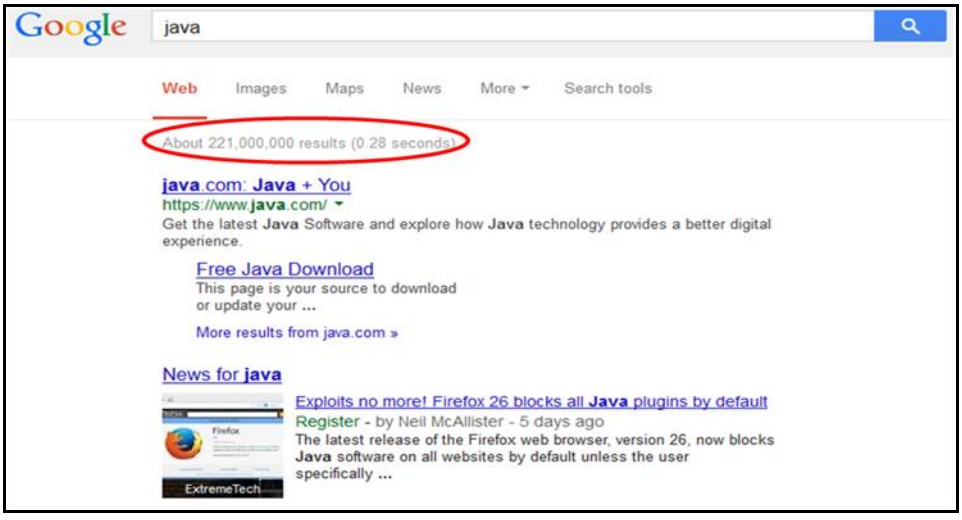

Figure 9. Results from Google using Keyword 'Java'

Figure 10. Results from Yahoo using Keyword 'Java'

The consequence of this huge number of results is difficulties for users to analyse every page of the provided outcome. Additionally, these results are mixed up in different categories. Therefore, SSE is proposed in this paper. The example of results for SSE using keyword 'Java' is shown in Figure 11. Three categories are listed they are Object-oriented programming language, Beverage and Island. The categories are based on the keyword entered. Each chosen category will produce results on the right side of the application related to the keyword and category.

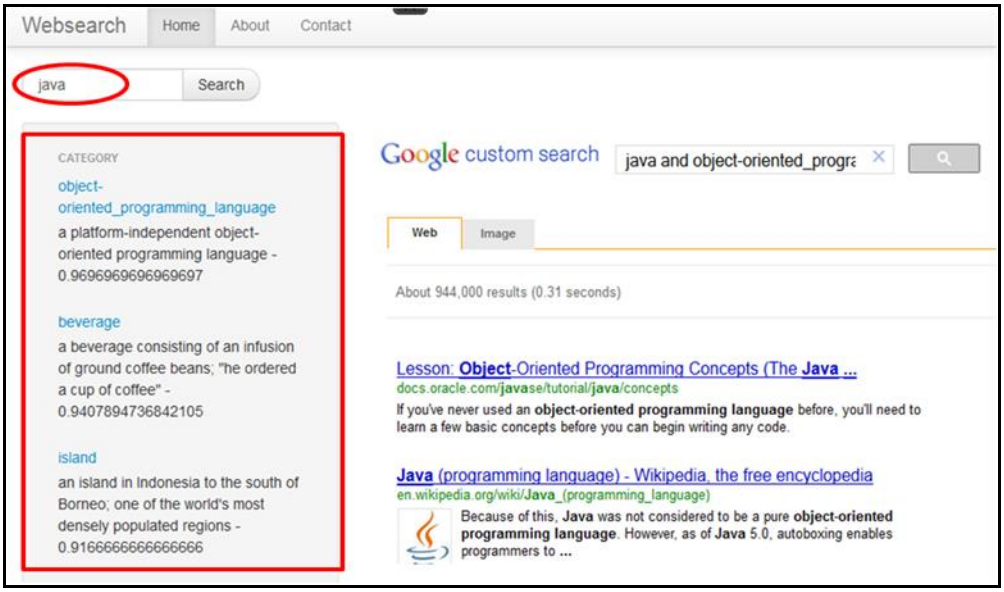

Figure 11. Categories for Keyword 'Java' using SSE

The results are listed in pages. Approximately, search engines give ten (10) links or web pages per page. It is shown in a rounded shape in Figure 12. For the comparison purposes, only ten (10) pages equivalent to almost 100 links are considered.

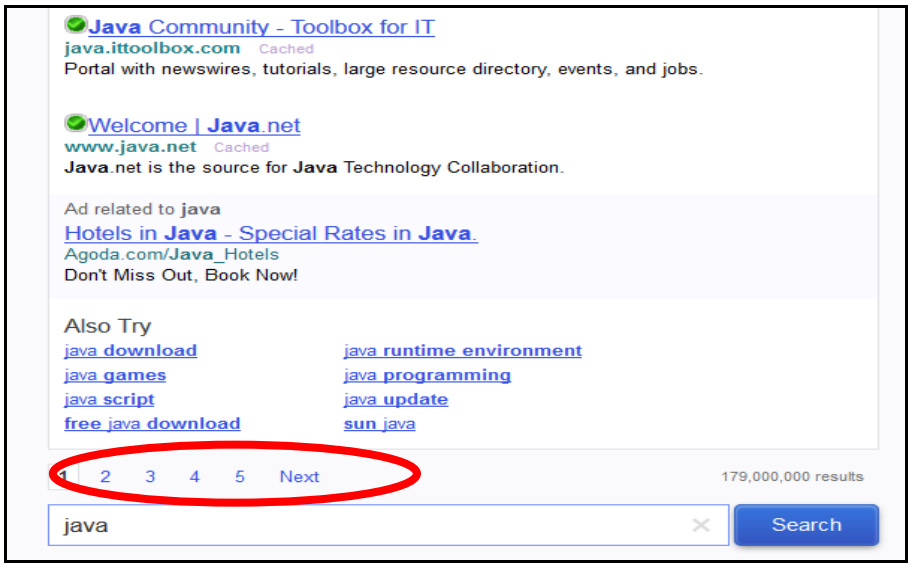

Figure 12. Pages Given by SSE in Providing Search Results

These data are compared with other search engine. In this stage, Google and Yahoo are used as comparison. Figure 13 shows SSE is giving full links that related to user's preferences. This is based on 100 first links given. Google only gives 84 links out of 100 links while Yahoo gives 90.

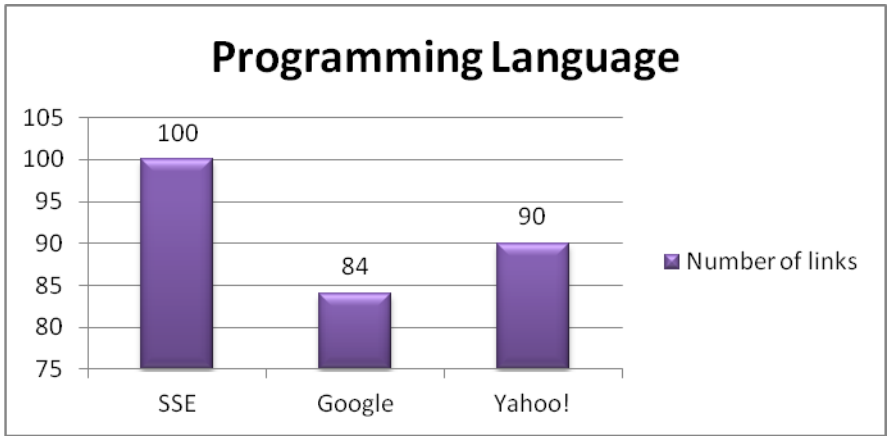

Figure 13. Results for Programming Language Category for three Different Search Engines

Table 3 listed number of web pages provided by the three search engines. SSE gives maximum number which is ten (10) links for every pages while Google and Yahoo give less than that.

Table 3. Results for Programming Language Category in every Page

Web search	Pages									
	1	2	3	4	5	6	7	8	9	10
	Links									
SSE	10	10	10	10	10	10	10	10	10	10
Google	7	10	10	9	9	5	10	8	8	8
Yahoo	8	8	8	9	10	9	10	10	10	8

The impact is not obvious from the Programming Language category. For example, Google uses Page Rank algorithm that listed popular web pages on the higher list of results. Java Programming Language is highly selected by internet users that make it popular based on that algorithm. Therefore, Google and Yahoo are still giving much number of links for this category. However Beverage category of Java in Figure 14 shows significant reduction in results acquisition. Google only gives six (6) results out of 100 first web pages while Yahoo only gives one (1) web page related to Java Beverage.

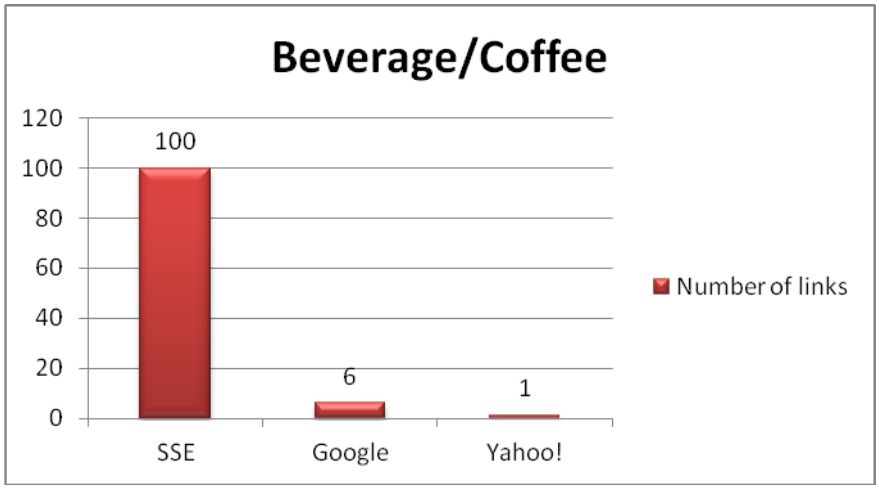

Figure 14. Results for Beverage Category for Three Different Search Engine

Table 4 shows results from Java category Beverage for each page. SSE gives better results when it comes towards personalization using user profiling. Ontology of the WordNet database gives more relevant information retrieval when using SSE.

Table 4. Results for Beverage Category in every Page

Web search	Pages									
	1	2	3	4	5	6	7	8	9	10
	Links									
SSE	10	10	10	10	10	10	10	10	10	10
Google	0	0	0	1	1	2	1	0	0	1
Yahoo	0	0	0	0	0	1	0	0	0	0

Based from Table 4, our semantic search engine (SSE) is compared using Google and Yahoo search engines. Results show that our SSE gives better information retrieval when it comes towards personalization using user profiling to produce more relevant results.

7. Conclusion and Future Work

WordNet usage has proven the capability in information retrieval especially regarding lexical issues. Various techniques can be done in manipulating the data. They are API or libraries, WordNet database and OWL ontology format. The attempts can assist in addressing issues in handling over increasing data and internet users. SSE is one of the alternative search engines for better information retrieval and managing information when it comes towards personalization using user profiling. The future work for this research is to test our SSE on the Cloud for its response time and usage as well as its relevant results.

Acknowledgements

This work is supported by Exploratory Research Grant Scheme (ERGS), Vote Number E006, Ministry of Higher Education, Malaysia and Universiti Tun Hussien Onn Malaysia (UTHM).

References

[1] Internet World Stats. Internet World Stats. http://www.internetworldstat.com, accessed **(2013)** September.
[2] C. D. Manning, P. Raghavan and H. Schütze, "Introduction to information retrieval", Cambridge: Cambridge University Press, vol. 1, **(2008)**.
[3] J. Mohamed Kassim and M. Rahmany, "Introduction to semantic search engine", In Electrical Engineering and Informatics, 2009, ICEEI'09. International Conference, IEEE, vol. 2, **(2009)**, pp. 380-386.
[4] A. Al-Maskari and M. Sanderson, "The effect of user characteristics on search effectiveness in information retrieval", Information Processing & Management, vol. 47, no. 5, **(2011)**, pp. 719-729.
[5] F. Shaikh, U. A. Siddiqui, I. Shahzadi, S. I. Jami and Z. A. Shaikh, "SWISE: Semantic Web based intelligent search engine", 2010 International Conference on Information and Emerging Technologies (ICIET), **(2010)**, pp. 1-5.
[6] D. Aguilar-Lopez, I. Lopez-Arevalo and V. Sosa-Sosa, "Toward the semantic search by using ontologies", In Electrical Engineering, Computing Science and Automatic Control, 2008. CCE 2008. 5th International Conference, IEEE, **(2008)**, pp. 328-333.
[7] SourceForge, JWNL (Java WordNet Library).http://sourceforge.net/projects/jwordnet, accessed **(2014)** September.
[8] SMU Lyle School of Engineering, Java API for WordNet Searching (JAWS). http://lyle.smu.edu/~tspell/jaws/index.html, accessed **(2013)** September.
[9] World Wide Web Consortium-W3C, WordNet RDF/OWL Files. http://www.w3.org/2006/03/wn/wn20/, accessed **(2014)** September.
[10] Stanford University, Protege. http://protege.stanford.edu/, accessed **(2014)** September.
[11] N. Abdullah, "Similarity Measurement in the Hybrid of Semantic Web Search Engine", International Journal of Computers & Technology, vol. 8, no. 3, **(2013)**, pp. 913-921.

