A HYBRID SEMANTIC SEARCH TECHNIQUE FOR WEB INFORMATION RETRIEVAL

NORYUSLIZA ABDULLAH

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

SEPTEMBER, 2015

ABSTRACT

Vast emergence of data on the web is an advantage in terms of availability. However, the ever-increasing growth of data and information makes finding the right information a challenge and an urgent task. This scenario results in the need to the improvement of information retrieval (IR). Web Information Retrieval (WIR) is the search engine has become the main resource in this area. Current WIR techniques have assisted in many ways, such as results ranking, categorization, and semantic searching. Nevertheless, there is a need to improve the current techniques to enhance information relevancy based on user's expectations. Therefore, in order to achieve the goals, a hybrid technique combining Categorization, Ontology, and User Profiling concepts is proposed in this research through the use of Semantic Web (SW) technologies. The objectives of this research were to design, implement and compare an alternative semantic search IR, and its effectiveness is tested in Cloud Computing (CC) environment. The WordNet, a lexical ontology resource, was used for keyword categorization as it consisted of large data in the English language, while the UTHM Ontology (UTHM Onto) supported User Profiling. The similarity between WordNet and UTHM Onto is generated using the semantic similarity measurement. The comparisons between the proposed Hybrid Search Engine (Hysse) with other techniques were identified based on Precision Effectiveness Metric. The term Java (referring to either a programme, beverage or an island) is used to measure the precision. The MAP of Java Object Oriented Programming Language for Hysse is 93%, WSP 89%, Doctopush 7%, Carrot2 73% and Google 93%. On the other hand, MAP of Java Beverage for Hysse is 81%, WSP 76%, Doctopush 9%, Carrot2 4% and Google 6%. Lastly MAP of Java Island for Hysse is 85%, WSP 82%, Doctopush 83%, Carrot2 3% and Google 11%. The Hysse is tested in CC using MYRENCloud and Amazon Elastic Compute Cloud (EC2). Comparison of Hysse and another technique which is Doctopush in cloud shows good results with the difference between them is only 14ms.

ABSTRAK

Kewujudan bilangan maklumat yang besar di laman sesawang memberikan kelebihan kepada pengguna. Walaubagaimanapun, ia menyebabkan proses pencarian maklumat menjadi lebih mencabar. Senario ini memerlukan penambahbaikan dalam proses *Information Retrieval* (IR). *Web Information Retrieval* (WIR) iaitu enjin carian telah menjadi sumber utama dalam bidang ini. Teknik WIR sediada membantu dalam beberapa aspek seperti menentukan kedudukan, pengkategorian dan carian semantik. Namun begitu, terdapat keperluan untuk memperbaiki teknik tersebut bagi memenuhi kehendak pengguna. Kajian ini menggabungkan konsep Kategori, Ontologi dan Profil pengguna dengan menggunapakai teknologi *Semantic Web* (SW). Objektif penyelidikan ini ialah merekabentuk, mengimplementasi dan membandingkan carian semantik IR yang ditambahbaik dan kebolehlaksanaannya diuji pada persekitaran *Cloud Computing* (CC). Wordnet sebagai sumber ontologi bahasa digunakan untuk mendapatkan kategori katakunci memandangkan ia mengandungi bilangan data Bahasa Inggeris yang besar. Manakala UTHM Ontologi (UTHM Onto) pula menyokong konsep Profil Pengguna. Pengukuran Persamaan Semantik digunakan bagi mengukur persamaan diantara Wordnet dan UTHM Onto. Ujian perbandingan diantara *Hybrid Search Engine* (Hysse) dengan teknik lain adalah berdasarkan berasaskan *Precision Effectiveness Metric* yang memberikan nilai *Mean Average Precision* (MAP). MAP bagi kategori *Java Object Oriented Programming Language* untuk Hysse adalah 93%, WSP 89%, Doctopush 7%, Carrot2 73% dan Google 93%. MAP bagi *Java Beverage* pula memberikan peratus untuk Hysse adalah 81%, WSP 76%, Doctopush 9%, Carrot2 4% dan Google 6%. Akhir sekali, MAP bagi *Java Island* untuk Hysse adalah 85%, WSP 82%, Doctopush 83%, Carrot2 3% dan Google 11%. Validasi Hysse dilakukan pada CC menggunakan MYRENCloud and Amazon Elastic Compute Cloud (EC2). Perbandingan Hysse dan teknik lain iaitu Doctopush pada *cloud* telah menunjukkan keputusan yang baik dengan perbezaan hanya 14ms.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 An Overview

Various sources existing on the web have given numerous information and knowledge. Search engines as the Web Information Retrieval (WIR) have become the main resource to capture those data. However, with the ever-increasing growth of data and information, retrieving the exact needed data and finding the right information on the search engines have become a challenge and an urgent task, whereby it often fails to give users their desired results. Furthermore, the use of natural language of information on the web is understandable by human, but difficult for computers to interpret (Ding *et al.*, 2005). The maturity of search engines should provide better mechanism to capture and obtain more information and making it meaningful for our purposes.

Current search engines are divided into two categories: Semantic Web (SW) and Semantic Search. SW does not propose a different architecture application, but instead, it gives information of a well-defined meaning and better cooperation between computers and people (Mikroyannidis, 2007). According to the inventor of the WWW and the father of the SW, this new idea is an extension to the traditional web that assists in expressing meaning (Berners-Lee, Hendler, & Lassila, 2001). SW pulls data from multiple sources and multiple formats. The Resource Description Framework (RDF) and Microformat features in SW allow websites to expose semi-structured information for machine use (Renaud, 2009). They deliver knowledge and assist in decision making. SW consists of languages and technologies that are intended to make the application development process and integration efforts a lot

simpler, faster, and more reliable (Wecel, 2003). On the other hand, Semantic Search is a process to obtain accurate results from typed keywords by analysing the intention of the information searcher. It involves an understanding of a term's context and meaning, and focuses on the text. Although they possess different meanings, both processes use SW technologies. A widely accepted technology in this field is ontology. Ontology is an explicit specification of a conceptualization (Gruber, 1993). More and more ontologies are produced and they are kept in ontology libraries for sharing and reusing (d'Aquin & Noy, 2012). Although ontologies are capable to help in producing good outcomes for the SW and the semantic search (Trillo *et al.*, 2011), researchers are trying to enhance the searching process in order to give better results by using the categorization/clustering technique (Carpineto *et al.*, 2009), and user profiling/personalization (Jie *et al.*, 2010; Antoniou *et al.*, 2010; Yoo, 2011; Moawad, 2012). However, there is an issue of relevancy that needs to be solved.

In addition to relevancy issue, cloud computing has been concerned in the WIR field. Most general search engines are implemented on cloud computing to manage huge data but not in semantic search engines. Cloud Computing is a large pool of easily usable and accessible virtualized resources, such as hardware, development platforms, and services. These resources can be dynamically reconfigured to adjust the variable load for optimizing resource utilization that is also known as auto-scaling (Luis et al., 2008). Cloud computing is not a completely new idea, but instead, it was initiated from time-sharing system in year 1960, and followed by network and grid computing in 1990 (Kim *et al.*, 2009). This new paradigm shifts the location of infrastructure away from desktops to the data centers to reduce the costs associated with the management of hardware and software resources (Brian *et al.*, 2008).

Recently, it seems to be the current trend among organizations and researchers to move towards cloud computing as it can give more advantages in the WIR due to its vast number of servers. Lee (2010) describes that if requirements become big and unpredictable, the on-demand nature of commodity resources become more attractive. The advantage of cloud computing technology has absolutely contributed to the optimization of information and knowledge retrieval since this process normally needs substantial amount of resources. Implementing the semantic search on cloud will help users in obtaining results effectively.

1.2 Research Problem

The World Wide Web (WWW) provides vast information to the end users. Nowadays, almost all information is available online. The evolution of Web 2.0 to Web 3.0 has encouraged the use of SW technologies in rendering better services (Bizer, Heath, & Berners-Lee, 2009). For that reason, the technologies have motivated to conduct this research. Search engines for the traditional web have helped in acquiring information on the WWW. However, users might experience headache when doing post-processing tasks due to the huge number of information.

As a matter of fact, irrelevant data has become one of the most critical issues in the searching process, and sometimes, hard to negotiate. Even the leading search engine like Google (Brin and Page, 1998; Franceschet, 2011) which is categorized as non-semantic search engine provides unrelated or useless result sets with reason that certain information is appropriate for specific users only, but not for all. Current IR strategies that concentrate solely on the keywords have failed to satisfy users' demands since they produce mixed-up outputs that needs manual analysis task.

Thus, new techniques were proposed in order to improve results acquisition and overcome the problem of information overload that causes irrelevant results (Jie et al., 2010; Antoniou et al., 2010; Yoo, 2011). Numerous researches have employed several different ways to address this matter including categorization (Carpineto et al., 2009) and categorization that based on lexical ontology (Trillo *et al.*, 2011). If not categorized, the search engine will give scattered results and the user would need to analyse every web page or link provided. This is a time consuming process and results of non-user-friendly environment. Numbers of studies have looked into categorization but studies in category ranking which follows user's profiling are rare to be found in the literature (Carpineto *et al.,* 2009).

Other researchers employ the concept of relevancy to present results that are only related to user's needs. In order to enhance the reliability and relevancy rate of search results, many researchers use User Profiling concept that analyse pages visited by users (Moawad et al., 2012). The User Profiling uses SW technologies that offer great potential (Esmaili, & Abolhassani, 2006; Lamberti *et al.*, 2009). The usage of SW has evoked several issues, such as the involvement of additional practice like folksonomies, limited scale of retrieval that works better in smaller data, and the chances of inaccurate information due to false ontologies (Hendler, 2007).

Another issue is handling the overwhelming data in search engine that is critical in this new era. For example Google as one of the largest internet search provider, processes more than 20 terabytes of raw web data per day (Rimal *et al.*, 2010). Convincing advantages offered by the cloud have helped in its penetration and adaptation for the WIR field. Furthermore, the existence of wireless networking, ubiquity of broadband, falling storage costs, and progressive improvements in internet computing software are the main contributors to the cloud computing emersion (Dikaiakos *et al.*, 2009).

Interesting features of cloud such as virtualization and pay-per-use concept have encouraged researchers to develop specific search engines to utilize the cloud services offered by the providers (Sheu *et al.*, 2009; Kang, & Sim, 2010). The idea of Cloud Computing has interest enterprises, countries and regions since the resources are scaled up or down based on the requirement and users only need to pay on-demand basis (Liu *et al.*, 2012). The service are obtained from the service providers, such as Amazon, Google, IBM, Citrix, and many more based on the service-level agreements (SLA) with vendors.

Search techniques that listed previously are divided into semantic and non-semantic. Non-semantic search engine such as Google and Yahoo! are utilizing cloud to handle their huge data. For example Yahoo! has built its private cloud called Sherpa to occupy a large-scale database system with the main objective is to build a flexible platform that corroborates rapid application changes and huge workload (Cooper, 2010). However, none of the semantic search engines are implemented in the cloud although they have possibility to process even larger data (Trillo *et al.*, 2011; Moawad *et al.*, 2012). Doctopush proposed by Trillo *et al.* (2011) and Web Search Personalizer by Moawad *et al.* (2012) are utilizing ontology to retrieve results. To compensate with the limitation, the proposed Hybrid Search Engine (Hysse) is suggested to be tested on Cloud Computing. Furthermore, the semantic search especially Hysse will be highly accessible through the utilization of light weight portable devices, for instance Personal Digital Assistant (PDA), iPad, iPhone, and smartphone since it is the aim of 21st century computing in accessing data and internet services (Dikaiakos *et al.*, 2009).

On the whole, the proposed hybrid technique of Hysse that is to be tested on the cloud is expected to give positive impact and to help boost research activities among students and academic staff. The need to obtain meaningful knowledge is

very prominent since it is the way to assist in conducting research. It may also assist in the teaching and learning processes. The outcome of this research is the hybrid of semantic search technique that gives more precise results. It has been tested on the cloud that assists in IR processes and the cloud platform gives shorter processing time regardless of number of users. The overall elements that bring Hysse up can be summarized in Figure 1.1. Listed are primary elements of the Hysse and they are utilized for the comparison purposes in Chapter 5.

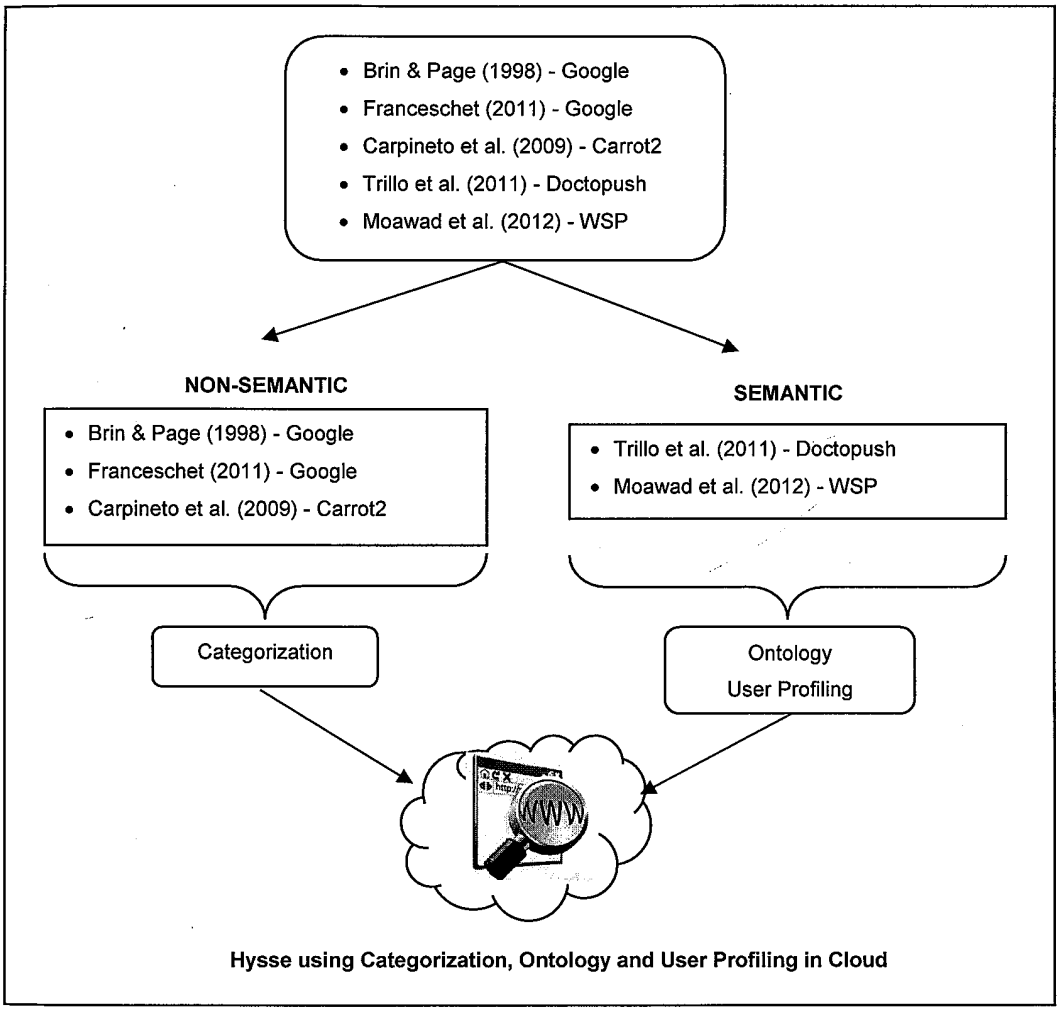

Figure 1.1: Primary Hysse Element

Based on Figure 1.1, Hysse is formed using combination of categorization technique from non-semantic search engine and ontology and user profiling techniques from semantic search engine. This combination of techniques is predicted to give more relevant results for information retrieval.

1.3 Objectives of the Study

The objectives of this research are:

(i) To design, implement and evaluate a Hysse technique using categorization, ontology and user profiling for web information retrieval.

(ii) To compare and validate the proposed Hysse technique in cloud computing environment with a different search technique based on loads and computational resources.

1.4 Scope and Limitations

This study focused on the data defined for Faculty of Computer Science and Information Technology - FSKTM, University Tun Hussein Onn Malaysia (UTHM). Keywords of Noun type are used in the testing phase instead of Verb or Adjective types. Single keyword is utilized throughout this research and the reason of this option will be explained on the next chapter. The Hysse is tested with other semantic and non-semantic search techniques based on a dataset of 500 Google's data. Precision Effectiveness Metric is utilized for comparison purposes in order to analyze the relevancy of results to the users. This research only utilizes two ontologies to do the categorization and ranking. The usage of more than two ontologies is highlighted in the future work in Chapter 6. A validation of Hysse's effectiveness on the cloud computing has been conducted using Stress Test analysis.

1.5 Research Significance

This research is to develop a systematic way in web IR using the Categorization, Ontology, and User Profiling concepts. After the breakthrough of using each method to solve the problem, the techniques were then combined as a complete unit. User's profile in ontology is studied to rank the most relevant category at the top to guarantee better choice of search results. This research is aimed to improve data search results in terms of information relevancy which will help research activities among students and academic staff in at least 20 public and 20 private institutions of

higher education in Malaysia. On the cloud computing side, this platform is used to validate the effectiveness of the proposed prototype. This study provides guideline in implementing cloud as a platform in semantic search engine. As a conclusion, this study strives to address the problem of retrieving relevance data effectively. It provides the solution to get faster and higher precision of results in a semantic search.

1.6 Thesis Outline

This thesis has six chapters. Chapter 2 discusses the literature review. It describes information and knowledge retrieval, web search engine, user profiling, and cloud computing. The subsection also includes non-semantic search and semantic search as techniques employed by search engines. The last two subsections list the discussions on cloud computing before the summary is presented in the last part of the chapter. Chapter 3 introduces the proposed technique - Hysse. It outlines the framework and phases of the prototype system. A brief explanation on the three phases involved is presented. This chapter also discusses the dataset and comparison method between the proposed prototype and other techniques. Chapter 4 discusses the implementation of Hysse in depth with all the involved procedures. The First Phase explains User Identification, followed by the Second Phase, which is Semantic Discovering. This phase portrays the WordNet and categorization based on the synset. Then, it discusses the core part of this research, which is User Profiling, including the development of UTHM Ontology, the integration of the ontology, the similarity measurement, the WordNet depth, the UTHM Onto depth, the Feature Matching, the Edge Counting and ranking. Chapter 5 explains the comparison between Hysse with other techniques. Statistical and graphical views represent analyses of the results. The validation of the prototype on the Cloud Computing environment is shown in this chapter. It also describes comparison between Hysse and other search technique in cloud platform. Chapter 6 summarizes the research, provides the list of contributions, and proposes potential future works.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter discusses the reviews of related literatures. The discussions start with Information Retrieval (IR) and Knowledge Retrieval (KR). IR and KR have been the foundation to modern web search engine. This chapter includes in-depth analysis on web search engines that are divided into non-semantic and semantic search. In KR, a user profiling concept and Lightweight Directory Access Protocol (LDAP) has been used as an enabler to implement that concept. It is a protocol that functions as a database which stores user information in Directory Information Tree (DIT). The last part discusses cloud computing since it has offered an interesting benefit to the web search and lastly, the cloud applications are presented.

2.2 Information Retrieval (IR)

Information Retrieval (IR) is an activity that involves information searching to satisfy users' need. This field has been conducted since 1950's due to the capability in handling information overload issues (Luhn, 1958). This starting point has triggered numerous improvements. Various applications have been developed in this area, including digital library, information filtering, media search, and search engines. The introduction of web search engines in 1990's has caused rapid development in this field. All IR applications, including web search engines that are also known as Web Information Retrieval (WIR) are actually based on traditional IR

model. Some of the models are Exact Match, Vector Space, and Probabilistic Approaches.

Along the way, huge improvements have been suggested by researchers and recent research trend in WIR is something related to SW. Several examples of the researches are those conducted by Shah *et al.,* (2002); Jiang, (2010); and Yang and Wu (2011). Shah *et al.* (2002) suggested text indexing and semantic markup to improve retrieval performance. In their research, they marked web pages semantically. The focus is different from the current web that is composed using natural language text. They used DAML+OIL languages and declared the language to be more powerful than RDF for knowledge representation.

Instead of focusing on marked web pages, Jiang (2010) has done his research on search engine as information retriever. He introduced Ranking Evaluator and Search Arbiter for rapid and correct information retrieval. Ranking Evaluator ranked the results to give more precise answer while Search Arbiter determined that the queries were processed by traditional keyword-based search engine or ontology-based search engine. On the other hand, Yang and Wu (2011) used lexical-based IR in their system that is WordNet. The issues of *synonymy* (same meaning or concept) and *polysemy* (more than one meaning) have been addressed. In that research, documents are annotated semantically using RDF/OWL. However, this process is no longer needed since the data is accessible from WordNet.

In addition, Trillo *et al.* (2011) and Moawad *et al.* (2012) have proposed interesting ideas using ontology in this field. Trillo *et al.* (2011) are concentrating on ontology and categorization while Maowad *et al.* (2012) are proposing ontology and user profiling concept. These researches displayed great results. Basically all the aforementioned researches have enhance the previous approaches such as the widely accepted PageRank algorithm by Google (Brin and Page, 1998, Franceschet, 2011) and Categorization (Carpineto *et al.*, 2009). The improved WIR concepts are utilizing the elements of ontology, RDF/RDFS, OWL, and SPARQL. The usage of ontology in current IR has successfully implements knowledge-based searching that supports Knowledge Retrieval and it is discussed in the next section.

2.3 Knowledge Retrieval (KR)

Knowledge is sitting at the higher level of data-information-knowledge hierarchy. Knowledge Retrieval (KR) is similar to Information Retrieval (IR) that has been mentioned in the previous section but with some modification. The aim of retrieving the knowledge is to improve the weaknesses that exist in the search process and data representation of Data Retrieval and IR. According to Yao *et al.* (2007), users are currently given an unprecedented amount of data and information. As a result, the ever-increasing growth of it has led to inaccurate knowledge. Data or information retrieval is inadequate in the current situation due to the lack of management at the knowledge level. Hence Yao *et al.* (2007) have proposed the Knowledge Retrieval Systems (KRS) to support knowledge discovery, organization, storage, and retrieval since there is a need to perform more tasks in addition to simple search. The KRS is focusing on semantics whereby knowledge is visualized in a structured way. The system extracted knowledge from information and convert it into a structure that human can use and organize the information for human usage.

In contrast, Tao *et al.* (2009) proposed an approach to rebuild user's knowledge systems through user's background knowledge from world knowledge base and user's Local Instance Repository. World knowledge is knowledge that is based on one's experience and education while Local Instance Repository collects data of visited web documents by the users.

The researches of Trillo *et al.* (2011) and Moawad *et al.* (2012) mentioned in the previous section although not directly state that their studies are categorized in KR but they also comprise KR characteristic. It is difficult to differentiate between IR and KR since KR with knowledge characteristic is embedded in the IR through the use of ontology. Yao *et al.* (2007) also stated that KR is a process of IR. To conclude the difference between IR and KR, KR is capable to give more meaning to data and information. It gives better results to the users. However, this research is categorized as IR with the characteristics of knowledge elements. The Web search engines that are used as tools in the IR are discussed in the next section.

2.4 Web Search Engines

Information availability is very limited in the past. The privilege in accessing them is allowed only to a certain group of people with high cost involved. With the existence of IR, the WWW and search engines have bridged the gap to collect and retrieve information. IR is a broad field and it did not begin with the web. Yet, WIR, which is also known as search engine, has become the primary tool to explore the internet. The specific purpose of the search engine is to search documents using keywords. Hence, in order to execute the search, crawler, spider or bot (robot) are used to fetch documents in WWW before they are extracted by users. This task needs to be conducted several times due to the frequently changing nature of the WWW. Indexer or catalogue indexes those documents based on words. These data are then utilized by the search engines to perform matching and ranking (Kassim & Rahmany, 2009).

While the process of document searching is proven successful in WIR, it is not sufficient to rely on keyword alone in finding the most related web site because substantial number of results obtained will increase time wasted in analysing the results. Hence, active researches on search engines have produced several forms of search techniques to meet users' needs and optimize the IR. Table 2.1 lists several commercial search engines by year, as presented by Seymour, Frantsvog, & Kumar (2011) and Kjuka (2015), with their features.

According to Manning, Raghavan, and Schütze (2008), the optimization of IR has contributed to the enhancement of web search engines that addresses information overload. This helps to implement the prominent task of distinguishing between right or wrong, or useful or useless information since currently the societies are dependent to the internet. The internet dependency in seeking information has been confirmed in Miniwatts Marketing Group report (2013) that claimed the growth of the internet users had been 566.4% from 2000 to 2012. Malaysia alone recorded 60.7% internet users until June, 30, 2012. This huge value proves that the internet is a good and accepted place to obtain information.

Table 2.1: Search engines (Seymour, Frantsvog & Kumar, 2011; Kjuka, 2015)

Search Engine	Year	Features
Archie	1990	Initial idea of internet searching. It uses file directories.
Gopher	1991	Menu system that distributes, searches, and retrieves documents over the internet.
Veronica	1991	Resource-Discovery system that searches file names and titles in Gopher index system.
Jughead	1991	Searches single server and indexes it. Slow in performance.
W3Catalog	1993	The first search engine. Mirrored pages on the web, reformats the contents, and implements dynamic querying.
Wanderer	1993	The first web robot that generates index.
Aliweb	1993	The second search engine, but it does not index site automatically.
Jump Station	1993	Uses web robot in finding web pages and indexing. Combines features of crawling, indexing, and searching, but limited to tiles and headings.
Web Crawler	1994	The first engine that provides full text search.
Meta Crawler	1995	Multiple search engines are used to generate search results.
Alta Vista	1995	A popular search engine, but shrinking with the existence of Google.
Excite	1995	Internet portal that uses a new crawler technique.
Dogpile	1996	Metasearch engine that searches multiple engines with page duplicate filtering.
Hotbot	1996	Updates database frequently to give updated results.
Inktomi	1996	Incorporates with HotBot search engine.
Ask Jeeves	1996	Answers questions, natural language, and keyword searching.
Northern Light	1997	Public and private custom search engine.
Google	1998	A popular search engine with PageRank algorithm.
Teoma	1999	Links popularity algorithm using specific subject.
Vivisimo	2000	A private enterprise search software company that sells search products.
Carrot2	2002	Categorization of results.
Yahoo! Search	2004	A popular search engine. Combines capabilities from acquired search engine companies, including Alta Vista.
MSN Search/Bing	2005	A search engine by Microsoft.
GoodSearch	2005	A Yahoo-powered search engine that donates revenue.
Wikiseek	2007	Indexed by Wikipedia pages and pages that are linked to Wikipedia articles.
Guruji	2007	An Indian internet search engine for Indian users.
Sproose	2007	A consumer search engine that allows users to vote for page ranking.
Blackle	2007	Aims at energy saving.
Powerset (Semantic)	2008	Natural language search engine.
Picollator	2008	Searches user visual query and/or text.
Viewzi	2008	Searches based on visual.

Table 2.1 (continued)

Search Engine	Year	Features
Cuil	2008	Organizes web pages by content with thumbnail pictures.
LeapFish	2008	A metasearch engine.
Forestle	2008	Inspired by ecology.
Valdo	2008	Focuses search for researchers in life sciences and biomedical, educators, students, clinicians, and reference librarians.
Goby	2009	A deep web search engine that searches selected database.
Exalead	2011	Searches software for users.
Cloudkite	2012	Search publicly shared files and reusable content.
Halalgoogling	2013	Islamic internet search that provides halal results.
JustCursor.com	2015	Minimalistic search engine.

Web search engines that listed in Table 2.1 are categorized as non-semantic and semantic search engine. Among all, only Ask Jeeves and Powerset are categorized as the semantic search type. Google as a non-semantic search has outperformed others. Non-semantic and semantic types search engines are explained in the next sections.

2.4.1 Non-semantic Search

Referring to Cloudio *et al.* (2009), WIR is divided into three methods. First, results are mixed in the list and users have to visit every link to find web pages that they need. This is known as general search engine. Second, categorization of the whole web in one group and the result is represented in one group for each category. Finally, the categories of the results are listed in a hierarchy of labelled clusters (parent-child) where every category has subcategories that provide details of the group. These results are produced by non-semantic search engine that is widely used and solely depends on keywords. Non-semantic search engines are divided into two categories which are General and Categorization. These categories are described next.

2.4.1.1 General Search Engine

Initially, the first type of search engine developed is the General Search Engine that is used to search for information in the WWW. Basically, it is keyword-based program that produces results from user's query. This type of search engine is more popular compared to others and has monopolized the WIR. For instance, the most popular search engines of this type are Google with the popularity percentage of 71.9% (Tumer, Shah, & Bitirim, 2009). It is the primary web search engine that was invented by Brin and Page (1998) with the PageRank algorithm that has become the success factor of Google (Brin & Page, 1998, Taneja & Gupta, 2010). It has given ultimate benefits to internet searchers since 1997. Numerous researchers have studied this algorithm including Franceschet (2011). Referring to Franceschet (2011), PageRank algorithm is recognizing the significance of a web page by looking into other important pages that are pointed to that web page. As an example the 'Java' keyword in Google search engine has produced 221 million search results. They are shown in Figure 2.1.

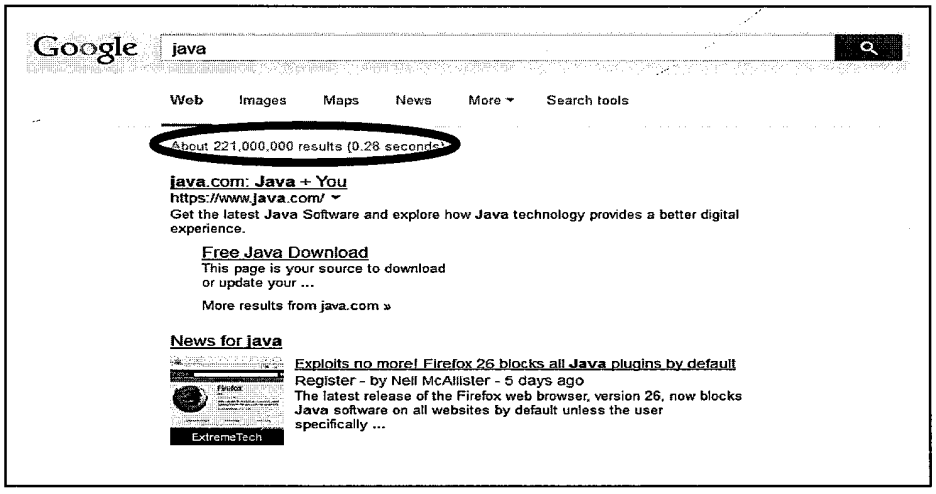

Figure 2.1: Number of Google's results using keyword 'Java'

This huge number of results from Google or any other general search engines would burden users to analyse every provided web pages or results. Additionally, these results are mixed up in different categories and listed in pages. It is shown in Figure 2.2. Approximately, the search engine gives ten links or web pages per page.

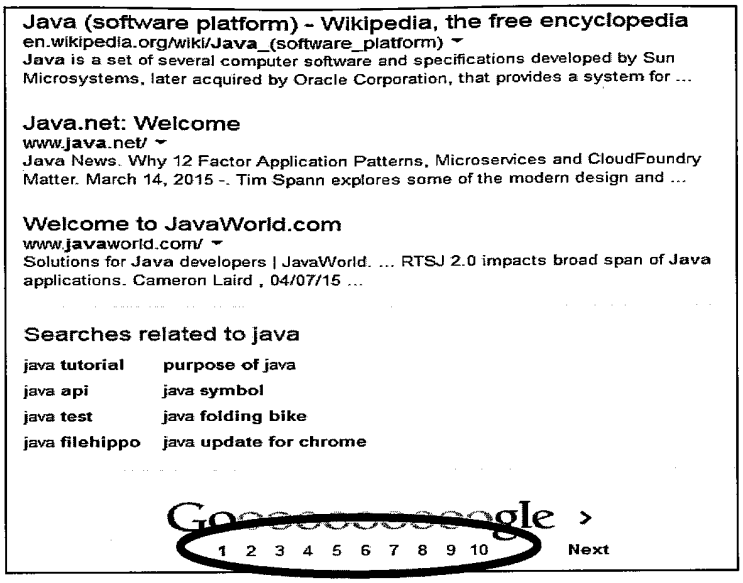

Figure 2.2: Pages given by the search engines in providing search results.

Google or any other general type of search engine produces loosely related keyword results. Hence, researchers are trying to reduce that drawback by proposing the categorization concept.

2.4.1.2 Categorization

Categorization is the process of organizing search results by topic and generally provides several categories that relate to the keyword. Category selection depends on user preferences. Sadaf and Alam (2012) state that it is a process of cumulating documents of similar groups to assist users in retrieving information faster by narrowing down the search result category. It is essential since information on the web is unstructured, disorganized, dynamic, heterogeneous, and huge. Some examples of the available search engines on the web that use this concept are Yippy (formerly known as Clusty), Carrot2, SnakeT, Kartoo, Grouper, and Open Directory. Among all, Carrot2 tool (Carpineto *et al.*, 2009) is studied in detail. Carrot2 is also known as Carrot Search. It is a type of search engine that provides topics of results in a categorized form. The example of Carrot2 results for 'Java' keyword is shown in Figure 2.3. Categories are listed on the left side of the search engines.

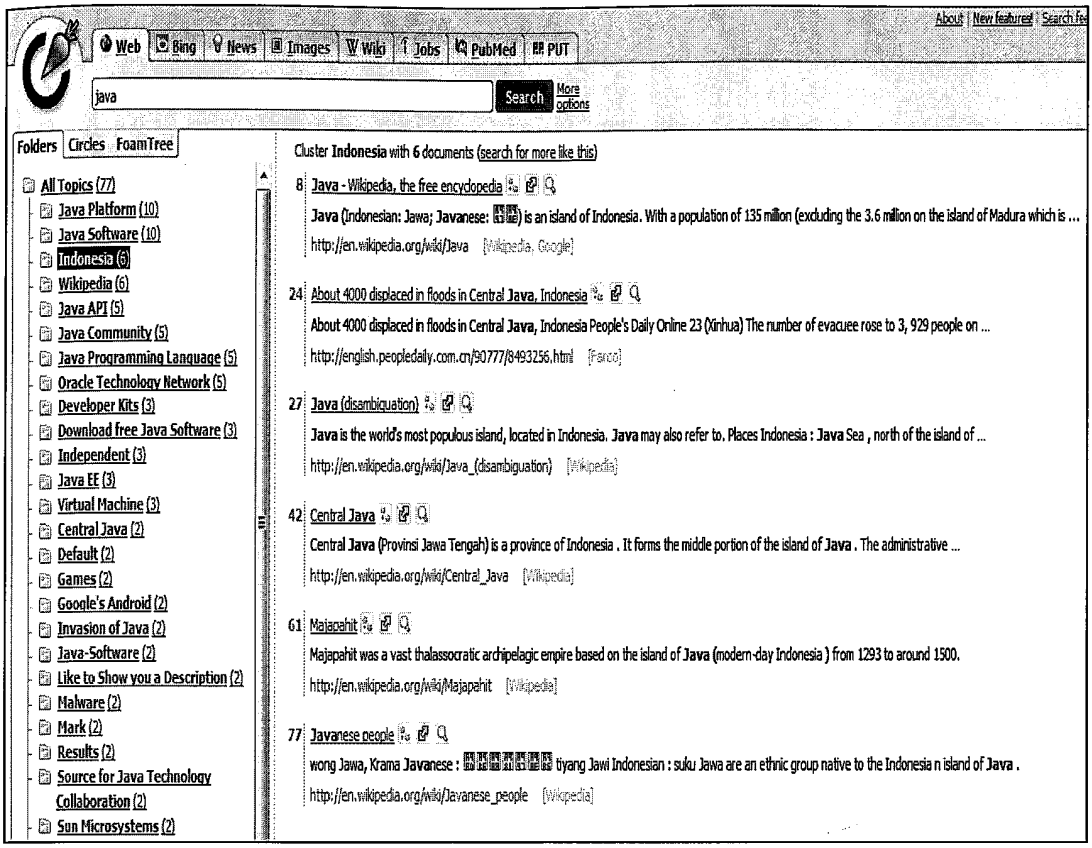

Figure 2.3: Carrot2 categorization search results

A huge number of categories are provided by the categorized search engines, which makes IR more complicated. By the time the data is obtained, Carrot2 gives 20 categories. Table 2.2 lists the categories for that search engine that uses categorization concept.

Table 2.2: 20 Categories of Java Keyword from Carrot2

Category ID	Category
C1	Java Development
C2	Java Platform
C3	Java Software
C4	Release
C5	Indonesia
C6	Wikipedia
C7	Java API
C8	Oracle Technology Network
C9	Java Programming Language
C10	Download free Java Software
C11	Independent
C12	Java Community
C13	Source Code
C14	Virtual Machine
C15	Central Java
C16	Default
C17	Google's Android
C18	Invasion of Java
C19	Java Sea
C20	Java Technology provides a Better Digital Experience

Some of the data on the table share the same domain, for example Java Development, Java Platform, Java Software and several other categories. These categories are separated although they are actually in the same group, which is Java Object-oriented Programming Language.

Although the categorization concept manages to provide better search results, it is inadequate and there is room for improvement. This is due to the insufficiency of

categorization alone to give optimum relevant results especially when there are many categories listed from the keyword. Hence, the idea of Semantic Search is proposed by researchers to give more relevant results using the science of meaning in language in order to meet user's demand. Furthermore, a good search engine requires assistance from the semantic or domain knowledge to answer intelligent queries (Shaikh *et al.*, 2010). Semantic Search is described in detail in the next subsection.

2.4.2 Semantic Search

Reliability and relevancy are two typical factors that usually affect the role of search engines (Shaikh *et al.*, 2010). Results are considered relevant if the retrieved resources are similar to what users have assumed (Lamberti, Sanna, & Demartini, 2009). Semantic search is critically needed since search effectiveness in IR depends on user's characteristics (Al-Maskari & Sanderson, 2011). According to the previous studies, users' search experience and high cognitive skills will produce 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 groups of users will benefit from the search engine. If no further action is taken, inexperienced and low cognitive skilled users will be left behind from receiving precise information within an acceptable time frame.

Furthermore, other research states that older internet users face difficulties in doing internet search, especially in navigation-oriented searching (Etcheverry *et al.*, 2012). Navigation-oriented reading can be defined as reading with the purpose to locate information, as opposed to content-oriented (reading to understand something). In other words, they are less efficient searchers compared to younger users. They usually perform internet search by following the same links and visiting the same pages. These cause increased time needed in executing tasks.

Although Etcheverry *et al.* (2012) have classified older users are those aged 65 and above, the statistics are related to the current trend of internet users since the aged population of the World is steadily increasing (Holzinger, Ziefle, & Röcker, 2010; Kinsella, & Velkoff, 2001). In Malaysian context, the formal retirement age has increased. In addition, more staff has prolonged their services, such as lecturers

and researchers, doing so on contract basis. This situation raises the number of older internet users.

These factors have encouraged researchers to perform research on the semantic search. One of them is Trillo *et al.* (2011). They integrate categorization with the semantic technique to group the output of searched keywords into different categories. In the studies they use semantic technology to define the possible categories. Referring to Andago, Phoebe, and Thanoun (2010), the semantic technology as the foundation of semantic web has been widely accepted to be incorporated in the semantic search engines to address the current approach's problems in searching for relevant information. The semantic technologies have become the backbone and are utilized in the semantic search, as further discussed in the next subsection.

2.4.2.1 Semantic Web (SW) Technologies

The internet and WWW have gone through tremendous growth. The evolution begins with Web 1.0, Web 2.0, and moving towards Web 3.0. In Web 1.0, WWW merely focused on static web pages designed for human readers, whereby information updates were managed by webmasters. This type of web requires human operator. The emergence of Web 2.0 changed the internet perspective to collective information among users with shared control that is focused on people and communications. Information is changeable by every individual. Web 2.0 has major features, including social networking sites, user created web sites, self-publishing platforms, tagging, and social bookmarking. These lead to new technologies, such as Facebook, Twitter, and YouTube, which encourage social interaction in an attractive and easy-to-use application.

The increase in users' needs and requirements promote attempts to improve in delivering a better usage of the WWW, especially when Web 2.0 has no longer fulfilled new requirements. Web 3.0, which is also known as SW and Web of Data, is expected to convey more in terms of users' needs. The SW helps in fetching information and other data related to it. Thus, it is not just sharing text of a page, but data and facts as well (Edwards, 2010). Referring to Jiang (2010), the benefit of using SW is the ability of the machine to understand descriptions of meaning.

Undoubtedly, the information explosion happening on the internet catalyses the existence of SW. Even though SW has different characteristics compared to the WWW, they are not separate entities, because this technology is an extension to the WWW. SW pages hold metadata that consist of definitions, notes, meanings, and many more. The major contributor to the success of the SW is ontology that is written in a structured and machine-readable form of RDF. RDF is a standard for describing web resources and understood by computers. Ontology has evolved and leads to many researches and development of prototype systems in SW. It enables accuracy in IR by exploiting a key content of SW resources (Esmaili, & Abolhassani, 2006; Lamberti *et al.*, 2009). Besides, SW delivers metadata, including schema and instances that describe things identified by Uniform Resource Identifier (URI) and are represented as graph structure.

The SW is generally confused with Semantic Search. Although overlap occurs between each of these terms, there is a clear distinction between them. SW is a web written in RDF to provide facilities in querying information from the web of data. This activity usually involves inference process. In contrast, the Semantic Search is a process of seeking information from normal web using the search engine to obtain more meaningful results to users compared to the traditional search technique. The similarity between them is the use of ontology. The benefit of ontology has encouraged most search engines to adopt this technique and produce semantic search, as discussed in the following section.

2.4.2.2 Ontology

The IR field has exploited Ontological concept in recent progress to improve knowledge searching and discovery mechanisms (Diez-Rodriguez, Murales-Luna, & Olmedo-Aguirre, 2008). Ontology is a new concept that points to the capability in sharing information structure among people, machines or applications, and currently has been adopted in the domain of reliable knowledge. This term has several interpretations. Derived from Greek philosophical study, it represents the nature of being, existence, or reality (Gruber, 1993). The expression is eventually widely used within computer science to describe the world consisting sets of types, properties, and relationships. Referring to widely cited Gruber's (1993) research paper, ontology is an explicit specification of a conceptualization. The ontologies need to specify

descriptions for classes in the domain of interest and relationships that can exist among things and properties that those things may have (Wecel, 2003). In a more descriptive form, ontology is a mechanism for representing formal and shared domain descriptions (Fluit, Sabou, & Harmelen, 2003), a vocabulary that can be used to express a knowledge base (Hepp, 2008), and functioned as a domain and knowledge representation (Edwards, 2010; Cardoso, 2007; Janev, 2010; Wecel, 2003).

As knowledge representation, it addresses the evolution of SW in Web 3.0 and it has been one of the semantic technologies utilized in Semantic Search. It has facilitated data representation to be more structured and easily interpreted by machine. Ontologies encode knowledge within a domain and also knowledge that spans within it. They include definitions of basic concepts in the domain and the relationships among them. SW needs ontologies with a significant degree of structure. The ontologies can contribute to express the contents of information and semantic relations between semantic elements. It can also support semantic reasoning, searching, and retrieval. Maier, Hadrich, and Peinl (2009) explain that documented knowledge, which spread across multiple sources such as web requires identification and visualization with the help of knowledge maps and integration supported by ontologies as a manager to semantic content.

Joo (2011) states research on ontology is necessary to ensure the diffusion of SW and semantic search. Hence, researchers have given great effort towards ontology and combining it with SW and semantic search since it is a promising research approach (D'Amato *et al.*, 2010). D'Amato *et al.* (2010) have manipulated the ontology using inductive reasoning that is capable to handle inconsistencies, noise and incompleteness of the knowledge bases. Certain tasks are required to manipulate and explore the ontology such as Find relevant resources, Select appropriate knowledge, Exploit heterogeneous knowledge sources, and Combine ontologies and resources (D'Aquin, 2008).

Generally, ontologies are divided into Upper, Domain and Application ontologies as shown in Figure 2.4. The upper ontology is a top-level or foundation ontology. De Bruijn (2003) and Diez-Rodriguez *et al.* (2008) describe upper ontology as a general concept and independent from particular task or domain. A number of upper ontologies have been developed, such as Base Formal Ontology (BFO), Cyc (or OpenCyc), Descriptive Ontology for Linguistic and Cognitive

Engineering (DOLCE), General Formal Ontology (GFO), PROTON (PROTo ONtology), Sowa's ontology, WordNet, and Suggested Upper Merged Ontology (SUMO) (Mascardi *et al.*, 2007). These ontologies facilitate capturing general and domain independent knowledge, for instance space and time (De Bruijn, 2003).

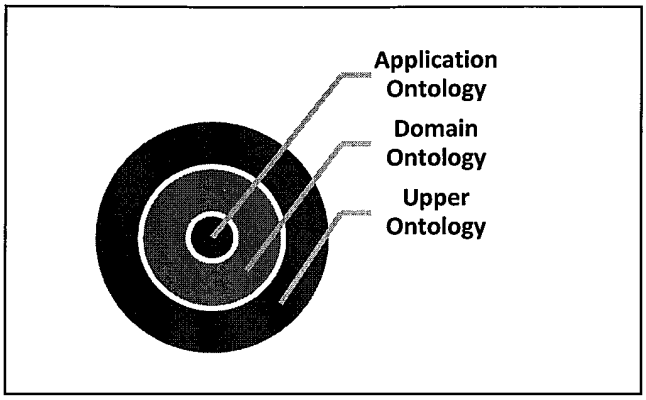

Figure 2.4: Types of ontology

However, the restriction of upper type ontology is the inability to satisfy user's need. Due to this, researchers are trying to produce domain ontologies. It models a specific domain, which represents part of the world. Various domain-specific ontologies have been developed, including domain in biomedical, pharmaceutical, university, currency, family relationship, and others. Although these ontologies are able to overcome the constraints experienced by the first type of ontology, something has to be done to resolve the specific requirements in application. Accordingly, application ontologies are created. Application ontology is a type of ontology that describes the concept depending on task. It represents scope, requirements and knowledge of the specific application. This ontology produces ontology classes to cater for a particular user (De Bruijn, 2003). Based on the criteria of the ontologies, integration between Upper, Domain and Application ontologies might give good results for WIR.

Although building an ontology is more complex in terms of logic and structure compared to building software (Cardoso, 2007), a promising bright future in its development and usage has encouraged researchers to develop any of the three ontologies (Trillo *et al.*, 2011 and Moawad *et al.*, 2012). Researchers and developers reuse the existing, well-established, and well-tested ontologies since it is more cost-effective than developing them from scratch. Ontology that can be reused greatly assisted in the expansion of this research field. In reusing ontologies, several steps

and techniques are implemented such as ontology assessment, integration, translation, and customization. The developed ontologies are available in ontology libraries. D'Aquin and Noy (2012) have listed 11 new generation ontology libraries: a)BioPortal, b)CupBoard, c)The OBO Foundry, d)oeGov, e)OLS, f)Ontology Design Patterns, g)OntoSelect, h)OntoSearch, i)The ONKI ontology server, j)The TONES repository, and k)Schema-Cache. An example of available ontology is Higher Education Reference Ontology (HERO). HERO is an ontology developed using the OWL API. It consists of higher educational structure.

The main goal of ontology engineering is to produce useful, consensual, rich, up-to-date, complete, and interoperable ontologies. Building ontologies that are linked to existing knowledge organization systems is needed to increase the interoperability between multiple representations or to increase access to the existing data (Hepp, 2008). In ontology engineering, several semantic markup languages for publishing and sharing ontologies on the WWW are utilized. They are used to describe the classes and relations between them (Wecel, 2003). The utilized languages are Web Ontology Language (OWL) and Resource Description Framework Schema (RDFS). For clarification, RDF is a framework for representing information in the web. Only a few things can be inferred from this language. Therefore, the RDFS was introduced. It is a semantic extension of RDF and is more expressive compared to the former. It stands between light-weight and heavy-weight ontologies. Dealing with heavy-weight ontologies needs more advanced language, Web Ontology Language (OWL).

In order to develop and edit the ontologies, several ontology editors can be utilized, including Protégé, OntoEdit, Topbraid, SWOOP, Neon Toolkit, Knoodl, OntoStudio, and Ontolingua. Protégé was chosen in this research to implement the UTHM Onto due to its support of a wide variety of plugins and imported formats. The modification of HERO is implemented using this editor. Querying and retrieving the ontology data are done using Simple Protocol and RDF Query Language (SPARQL).

As highlighted before, semantic search is utilizing several types of ontologies. WordNet as one of the Upper type ontology known as Lexical ontology has been widely used in the WIR. It is discussed in the next section.

2.4.2.3 WordNet

Wordnet is a machine-readable dictionary (Miller, 1995). It is one of the Upper Ontology resources that is used to interpret the natural language properly. Many Upper type ontology resources are available, as shown in Table 2.3. Among all, the most related product with WordNet is CYC. WordNet is used due to its suitability in finding similar English words through lexical resource. English language is emphasized because a large number of internet resources are written in English language. IR using other languages, such as Malay, Japanese, German, and France, are possible to implement, nevertheless, it is out of the scope of this research.

Wordnet has become an ideal tool for disambiguation of meaning, semantic tagging, and IR based on its design that can be easily manipulated by computers and free of charge (Morato *et al.*, 2004), and provides huge data. Moreover, WordNet is accessible, offered good quality of ontology and have high potential in processing the natural language. Hence, many researchers are utilized this ontology. Trillo *et al.* (2011) has integrated WordNet together with other ontologies in their research. Although it gives better results but it uses more processing time.

Table 2.3: Upper Type Ontologies

Ontology	Characteristic
Wordnet	Lexical reference system.
CYC	Provides knowledge based on everyday common sense knowledge.
Basic Formal Ontology (BFO)	Supports domain ontologies developed for scientific research.
General Formal Ontology (GFO)	Specializes in persistence and time model.
EuroWordNet	Consists of European Language interconnected with Interlingual Index (ILI).
SUMO	Upper ontology of computer information processing systems.
DOLCE	Supports natural language and human common sense.

WordNet groups English words into sets of synonyms, antonyms, and homonyms (hypernym/hyponym relationships), as illustrated in Figure 2.5, and records various semantic relations between them. From the illustration, *Colour* is the

REFERENCES

Ali, R., & Beg, M. M. (2011). An overview of Web search evaluation methods. *Computers & Electrical Engineering*, 37(6), 835-848. Elsevier.

Al-Maskari, A., & Sanderson, M, (2011). The effect of user characteristics on search effectiveness in information retrieval, Information Processing & Management, 47(5), pp. 719-729.

Andago, M. O., Phoebe, T., & Thanoun, B. A. M. (2010). Evaluation of a semantic search engine against a keyword search engine using first 20 precision. *Intern. Journal for the Advanc. of Science & Arts,* 1(2), 55-63.

Antoniou, D., Paschou, M., Sourla, E. and Tsakalidis, A. (2010). A Semantic Web *Personalizing Technique: The Case of Bursts in Web Visits.* Proceedings of the Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on. pp. 530-535.

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I. & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM, 53*(4), 50-58.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), 1-22.

Brian, H. A. Y. E. S., Brunschwiler, T., Dill, H., Christ, H., Falsafi, B., Fischer, M., & Zollinger, M. (2008). Cloud computing. Communications of the ACM, 51(7), 9-11.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer networks and ISDN systems, 30(1), 107-117. Elsevier.