Introduction to Semantic Search Engine

Junaidah Mohamed Kassim and Mahathir Rahmany
Faculty of Technology and Information Science
University Kebangsaan Malaysia
junaidah@ftsm.ukm.my and Mahathir.rahmany@yahoo.com

Abstract — Search engine is the most important tool to discover the any information in World Wide Web. In a row with the terrific growth numbers of the web, traditional search engine nowadays is not appropriate anymore to be used. Searched by keyword and do not understand polysemy and synonymy are some reasons the traditional search engine is not suitable anymore. Semantic Search Engine is born of traditional search engine overcome the problem below. This paper will discuss about semantic search engine and the traditional search engine.
Keywords: Semantic Search Engine, Information Retrieval.

I. INTRODUCTION

Search engine has become a primary need to explore the internet. Without Search Engine, there are no uses of information in website, blog, etc; because without search engine, it is almost impossible to look for one by one website just for search information in internet.

In a row with the extraordinary growth of web, there are many search engines come out to help the users on finding their need, but search engines find it increasingly difficult to provide useful results. To manage this explosively large number of web documents, automatic clustering of documents and organising them into domain dependent directories became very popular. [1]

The terrific increment of the web has made the evolution of the web itself. From web 1.0 (first generation of internet – 1990 – 2000), web 2.0 and now has become to web 3.0.

Web 1.0 refers to Internet at its emerging stage, with corporate and institutional websites occupying 90% of the cyber space, with an one-way mode. Thus Internet access serves functional purpose, and people could also read and extract information from the websites. Internet access was achieved typically through telephone dial-up at that time [2]. “Web 2.0” is transforming the Web into a space that allows anyone to create and share information online—a space for collaboration, conversation, and interaction; a space that is highly dynamic, flexible, and adaptable [3].

Web 3.0 is one of the terms used to describe the evolutionary stage of the Web that follows Web 2.0. Given that technical and social possibilities identified in this latter term are yet to be fully realized the nature of defining Web 3.0 is highly speculative. In general it refers to aspects of the Internet which, though potentially possible, are not technically or practically feasible at this time. [4]. We can see a differences between web 2.0 and 3.0 in figure 2.

From the figure 1, it shown that web 1.0 is a one-way platform, web 2.0 is a two-way platform where

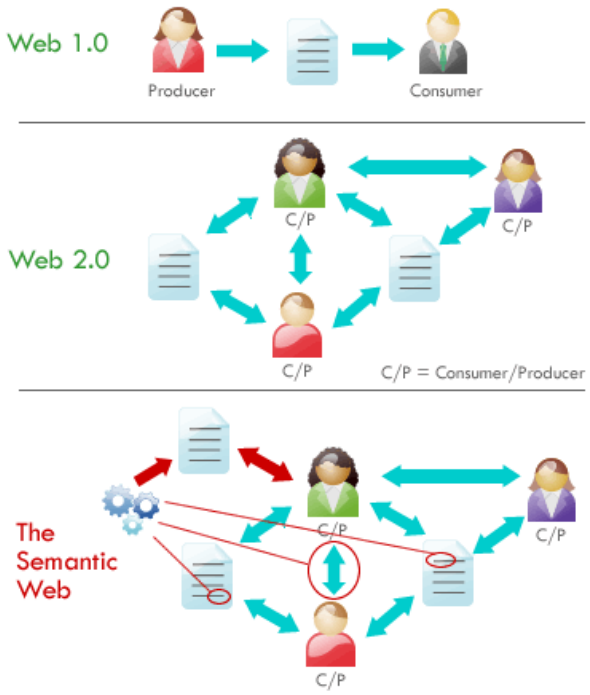

Figure 1: Evolution in web 1.0, web 2.0 and web 3.0 [5]

participation is a key-word. While the Web 3.0 shows more intelligences: the "web machine" learns, suggests and anticipates what people like and would like to get [5].

II. LITERATURE REVIEW

A. Information Retrieval

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that

978-1-4244-4913-2/09/$25.00 ©2009 IEEE

Authorized licensed use limited to: VIT-Amaravathi campus. Downloaded on September 25,2024 at 07:13:22 UTC from IEEE Xplore. Restrictions apply.

satisfies an information need from within large collections (usually stored on computers). [7]

Figure 2: Differences between web 2.0 and web 3.0 [6]

	Web 2.0	Web 3.0
Main task	Focus the power of community to create dynamic contents and interaction technology	Linked data, devices and people across the web
Linking	Walled gardens inhibit interoperability	Data and devices linked more easily and in new ways
Content	Individual and organization create content	Individual, organization, machine create content which can be reused
Technology	AJAX	Resource Description framework (RDF)
Website	Google, facebook, wikipedia, ebay, youtube	Dbpedia, sioc-project.org

The aim of an information retrieval system is to find relevant documents thus relevance is a (if not 'the') central concept of information retrieval. [8]

From the figure 1 show how an information retrieval system work in generally. It starts with general information or query from the user. Then the query will be extracted to be metadata. This process usually is called creating representation.

In the other hand, corpus (corpus is a collection of one or more documents, typically related, and available to an information retrieval system) is also extracted from formatted document to full text index. This process is done by several steps, such as Remove properties and formatting, parser, stopword removal, stemming and Synonym matching. That full text index will be saved in index objects database.

Next step is performed comparison between user query and index object to retrieve objects. Then last, Evaluate the result or refine search if the result is not hesitated.

B. Search Engine

Search engine is one of the most famous tools to discover the information in internet. General purpose search engines have typically used exhaustive crawlers to build and update large collections of documents. [10]

All search engines consist of three parts: (1) a database of web documents, (2) a search engine operating on that database, and (3) a series of programs that determine how search results are displayed [11-6]. Part of the search engines' success might be due to their simplicity: you enter some words and the results are then output in form of a ranked list, in which the search engine estimates the relevance of each indexed website to your query. [12]

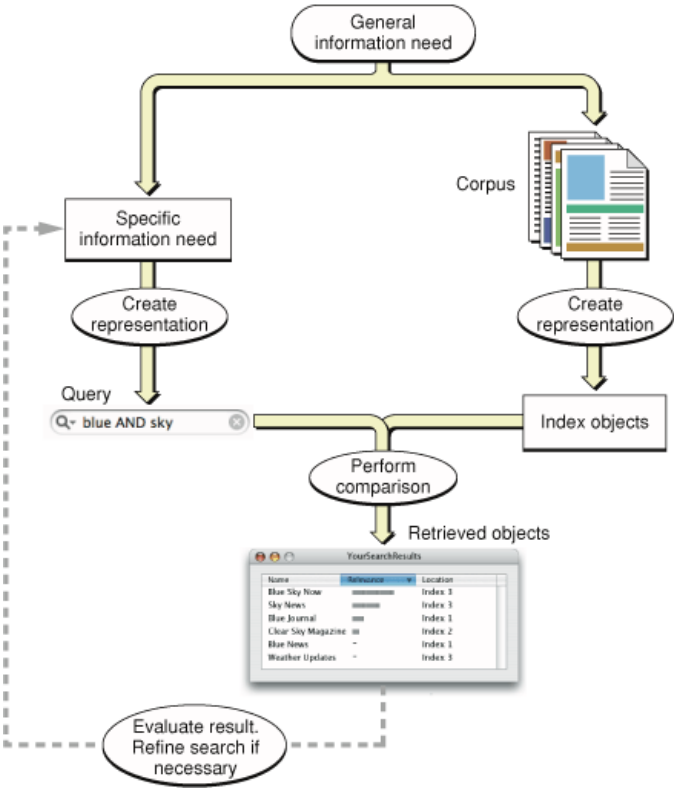

Fig. 3 An information retrieval system [9]

The utility of any search engine depends on two parts: the quality of the system, and content, which in our case is provided by a large number of contributors (personal and corporate web sites, for example). Importantly, content suppliers have to agree on a social contract (as anywhere on the Internet) on how to provide and publish data [13].

Now days, as one of essential tool in internet, search engine has many types, such as Crawler-Based Search Engines, Directories, Hybrid Search Engines, Meta Search Engines and Specialty Search Engines.

Crawler-based search engines use automated software programs to survey and categorise web pages. The programs used by the search engines to access your web pages are called 'spiders', 'crawlers', 'robots' or 'bots'. A 'directory' uses human editors who decide what category the site belongs to; they place websites within specific categories in the 'directories' database. Meta search engines take the results from all the other search engines results, and combine them into one large listing; and Specialty search engines is the search engine for

Authorized licensed use limited to: VIT-Amaravathi campus. Downloaded on September 25,2024 at 07:13:22 UTC from IEEE Xplore. Restrictions apply.

searching in niche areas, such as shopping, local search, etc. [14]

There are several components in search engine, such as crawler, indexer, sorter, analyser, and searcher.

Crawler is a program that is created to visit each webpage periodically and collect the important information and store them into database. The Web Crawler Application is divided into three main modules.

a. Controller Module - This module focuses on the Graphical User Interface (GUI) designed for the web crawler and is responsible for controlling the operations of the crawler. The GUI enables the user to enter the start URL, enter the maximum number of URL's to crawl, view the URL's that are being fetched. It controls the Fetcher and Parser.
b. Fetcher Module - This module starts by fetching the page according to the start URL specified by the user. The fetcher module also retrieves all the links in a particular page and continues doing that until the maximum number of URL's is reached.
c. Parser Module - This module parses the URL's fetched by the Fetcher module and saves the contents of those pages to the disk. [15]

After that indexer create index in the database to organize the data by categorize them. The indexer extracts all the information from each and every document and stores it in a database. All high-quality search engines index each and every word in the documents and give a unique word Id. Then the word occurrences, which some search engines call "hits," are checked, recording all the words, including their placement in the document, their font size and capitalization. [16]

C. Directory

A directory is "human-powered". Humans select sites that will be listed. Directories list sites under categories and subcategories, becoming more specific at each level. Directories include a fraction of the sites available via a search engine. Many search engines now include a directory feature. Portals are directories for very specific topics and audiences [17].

Web Directories can be real time savers. Why go through the trouble of searching and then scrolling through several pages of links when someone else has already selected the best links on a subject for you? [18].

Every Open Directory Project (ODP) directory has an associated URL, which contains a description of the directory and a number of Web sites that have been manually listed as pertaining to the directory topic, accompanied by brief descriptions of each site. This information is completed with a list of subdirectories, each containing more Web sites and subdirectories. Finally, some directories also have pointers to the same category in other languages [19].

One of the famous directory is DMOZ (Directory Mozilla). In 1998, Rich Skrenta began the volunteer-run GnuHoo (name inspired by GNU freeware group), which would later become the ODP. On June 5, 1998 GnuHoo went live and within two weeks "there were 200 editors, 27,000 sites, 2,000 categories". Within five months the site garnered the attention of Netscape and was acquired for their Netcenter site. However, the site's name had now been changed from GnuHoo to NewHoo due to naming conflicts with the freeware group called GNU who was in the process of developing a free UNIX inspired operating system. Soon the name changed one final time, becoming the Open Directory Project [20]. Now DMOZ has 4,598,551 sites - 82,103 editors - over 590,000 categories.

D. Semantic Search Engine

Differently from traditional search engines, a semantic search engine stores semantic information about Web resources and is able to solve complex queries, considering as well the context where the Web resource is targeted. [21]

Semantic search integrates the technologies of Semantic Web and search engine to improve the search results gained by current search engines and evolves to next generation of search engines built on Semantic Web. [22]

In general, processes of semantic search engine are: (1) The user question is interpreted, extracting the relevant concepts from the sentence, (2) that set of concepts is used to build a query that is launched against the ontology, and (3) The results are presented to the user. [23]

E. Ontology

According to T. R. Gruber, an ontology can be defined as an explicit specification of a conceptualization [24]

The term of ontology is come from philosophy, that is "*the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality*". [25]

In the internet, we can see the example of the ontology, such as Yahoo! Categories, DMOZ Directory, Amozon.com product catalogue, WordNet, GO (Gene Ontology) (www.geneontology.org), Unified Medical Language System (UMLS) and UNSPSC - terminology for products and services.

Kinds of ontology that are various commonly used are (a) Terminological ontologies where concepts are word senses and instances are words, (b) Topic ontologies where concepts are topics and instances are documents, and (c) Data-model ontologies where concepts are tables in a data base and instances are data records (such as in a database schema). (26)

There are several reasons why someone wants to develop ontology, such as to share common understanding of the structure of information among people or software

Authorized licensed use limited to: VIT-Amaravathi campus. Downloaded on September 25,2024 at 07:13:22 UTC from IEEE Xplore. Restrictions apply.

agents, to enable reuse of domain knowledge, to make domain assumption explicit, to separate domain knowledge from the operational knowledge and to analyze domain knowledge. [27]

Ontology describes knowledge about the domain in terms of concepts or vocabularies within the domain and relationships between them.

Ontology is needed to develop semantic search engine. By using semantically richer ontologies the following benefits can be obtained. Firstly, ontologies can be used to describe the domain knowledge and the terminology of the application in more detail. For example, relations between categories in different views can be defined. Secondly, ontologies can be used for creating semantically more accurate annotations in terms of the domain knowledge. Thirdly, with the help of ontologies, the user can express the queries more precisely and unambiguously, which leads to better precision and recall rates. Fourthly, through ontological class definitions and inference mechanisms, such as property inheritance, instance-level metadata can be enriched semantically [28].

III. SEMANTIC SEARCH ENGINE

A. The Architecture of Semantic Search Engine

A typical semantic search engine should consist of the following components: (1) Ontology development, (2) Ontology Crawler, (3) Ontology Annotator, (4) Web crawler, (5) Performing semantic search, (6) Query builder, and (7) Query pre-processor. [29]

Some of the reason why someone want to develop an ontology are (1) To share common understanding of the structure of information among people or software agents, (2) To enable reuse of domain knowledge, (3) To make domain assumptions explicit, (4) To separate domain knowledge from the operational knowledge, and (5) To analyze domain knowledge. [30]

An ontology crawler can crawl through the Web to find new ontologies and populate the database with them. Note that the ontologies crawled by the ontology crawler will not be directly dumped into the database. Ontology translator, with the help of ontology mapper, will translate them into the database tables. [31]

Ontological annotations identify real-world entities alongside properties and relations that characterize the entities' attributes and role in their textual context, with respect to a reference ontology. [32]

Web crawler is an automatic program (sometimes called a "robot") which explores the World Wide Web, following the links and searching for information or building a database - such programs are often used to build automated indexes for the Web, allowing users to do keyword searches for Web documents - it is also the name used by one of the most popular publically available keyword searching engines. [33]

(A) Ontologies are created in plain text format (.OWL or .DAML). Ontology translator translates them into relational database tables. Ontology Crawler finds new ontologies on the Web and adds them to the ontology library. (B) Users use Ontology Annotator to annotate their Web pages with these ontologies and publish them on Web. (C) Web Crawler crawls the Web to find Web pages annotated with these ontologies and builds knowledge base from the instances of these ontologies in these Web pages. (D) Users construct search queries with the help of Query Builder and these queries are sent to the Inference Engine after pre-processing by Query Pre-processor. (E) Inference Engine carries out reasoning on these search queries by using ontology database and the knowledge base. The results are finally sent to the user to be viewed on the Web.

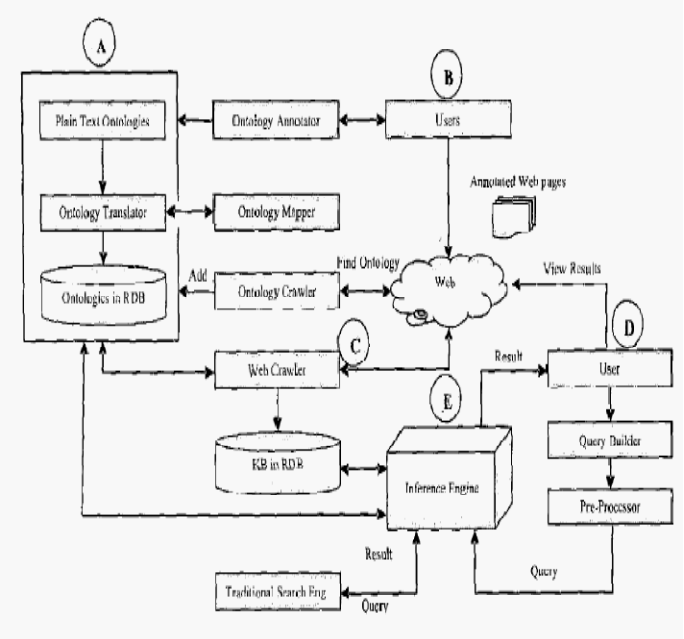

Figure 4: Proposed layout of semantic search engine

B. The Use of Semantic Search Engine

Semantic search attempts to improve the results of research searches in two ways, a) traditional search results take form of a list of a document/web pages and b) the search phrase in research searches typically denotes one (or occasionally two) real-world concepts. [34].

There are several advantages of semantic search, such as

- A semantic search makes it easier to locate relevant information to the user's subject of interest, saving the user a lot of time reading unrelated Web pages. [35]
- A semantic search engine can handle of long-tail queries. Without relying on statistics, long-tail queries can be analyzed by semantic algorithms

Authorized licensed use limited to: VIT-Amaravathi campus. Downloaded on September 25,2024 at 07:13:22 UTC from IEEE Xplore. Restrictions apply.

on the fly, and bring search results with the accurate context. [36]

C. Why Semantic Search Engine is Important

Finding what we need is often a hard job. Current search engine technology is very good in finding complete Web pages published all over the world, but it lacks the desired precision and recall when searching for multimedia resources. [37].

Semantic searches can overcome the aforementioned limitations of keyword searches because they use an ontology to infer information about objects. This enables a semantic search system to correctly identify objects even when the object's associated metadata does not explicitly match the user's search criteria. The ability to infer information based upon relationships encoded in the ontology enabled users to identify objects which are logically related. For example a database may contain objects from multiple domains whereby researchers in each discipline use there own terminology. To locate all desired objects using a keyword search would require the user to have knowledge of the relevant terminology used by workers in each discipline. This type of information can be encoded in an ontology and the semantic search system could automatically retrieve the semantically related items. [38]

IV. DISCUSSION

A. Traditional Search Engine vs. Semantic Search Engine

There are some differences between traditional search engine and semantic search engine. The differences are:

Traditional Search Engine:

- In engine prompt, you are entering keyword.
- Do not understand polysemy and synonymy.
- Unknowing meaning of the terms.
- Do not take into account stop words such as a, and, is, on, of, or, the, was, with, by, after, the.
- When looking at a web page, a conventional search engine looks for the distribution of words within the web page to try and find how relevant it is to the user's search query. Basically, this means that a web page with similar words to those the user types into a search engine will be thought to be more relevant, and will appear at a higher position in the search results page. [39]
- Unable handle long-tail queries.

Semantic Search Engine:

- In engine prompt, you are entering question.
- Understand polysemy and synonymy.
- Knowing meaning of the terms.
- Take into account stop words such as a, and, is, on, of, or, the, was, with, by, after, the.
- Is designed to try and understand the context that the words are used within the web page to try and match it more accurately to the user's search query. [40]
- Able handle long-tail queries.

B. Comparison between Traditional Search Engines vs. Semantic Search Engine.

For the comparison between Traditional Search Engines vs. Semantic Search Engine, please see these two search engines. Hakia (www.hakia.com) as a semantic search engine and Dogpile (www.dogpile.com) as a traditional search engine.

The keyword phrase that is used for this comparison is “what is the weather in Kuala Lumpur”. See the result below.

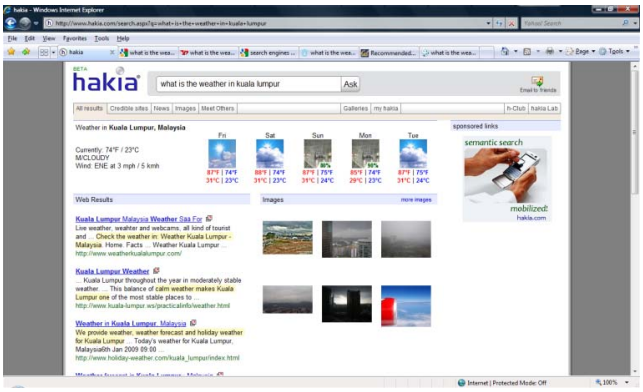

Figure 4: Hakia

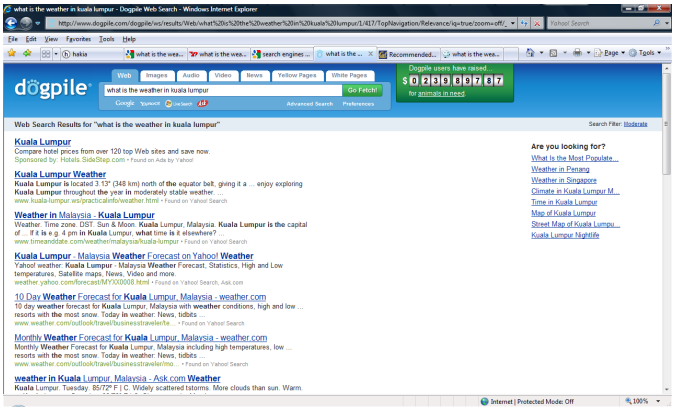

Figure 5: Dogpile

From the figures below is showed that Hakia seem knows what their user want. It is clear from the result that showed by Hakia. They show the information of Kuala Lumpur weather not the website that contain keyword of Kuala Lumpur weather like Dogpile.

V. SUMMARY

Authorized licensed use limited to: VIT-Amaravathi campus. Downloaded on September 25,2024 at 07:13:22 UTC from IEEE Xplore. Restrictions apply.

Analogously with the arise of web site that is directly proportional of the internet user growth, has made the traditional search engine not able to provide the precise result. It is because the weakness of that search engine which search just based on keyword and also the traditional search engine does not care about polysemy and synonymy. Sometimes, when we are using the traditional search engine the result that is showed is not satisfy to the user needs.

Semantic search engine is became an answer to overcome the lack of traditional search engine. It is not like traditional search engine which search based on keyword, semantic search engine try to analyze and understand the user want by doing logical reasoning, so the result will more precise. Semantic search engine also can handle polysemy, synonymy and long-tail queries well. It is no doubt if semantic search engine will show the result almost that you want.

One of the advantages of semantic search engine is no need to open the website one by one just to check the content, because the search engine will give you the precise result which is not the wor.

VI. REFERENCES

[1] Debnath, Sandip., et al. (----). Knowledge Discovery in Web-Directories: Finding Term-Relations to Build a Business Ontology, The Pennsylvania State University, University Park, USA.

[2] The Development of '.hk' with Internet Trends From Web 1.0 to Web 3.0. Available: https://www.hkdnr.hk/webupdate/article/pdfs/RTHK%20Article%20Final%20_Eng_%20revised.pdf

[3] Building a Library Web Site on the Pillars of Web 2.0. Available: http://www.infotoday.com/cilmag/jan07/Coombs.shtml

[4] Web 3.0. Available: http://en.wikipedia.org/wiki/Web_3.0

[5] From web1.0 to web3.0: get the point in a picture. Available: http://fredericmartin.typepad.com/myblog/2007/11/from-web10-to-w.html

[6] Web 3.0. Available: http://webuser.hs-furtwangen.de/~heindl/ebte-08ss-web-20-Suphakorntanakit.pdf

[7] D. Manning, Christopher., et al (2008). Introduction to Information Retrieval. cambridge university press, New York.

[8] Mizzaro, Stefano (----). How many relevances in information retrieval. University of Udine, Italy.

[9] ---, (2005). Search Kit Programming Guide. Apple Inc, Canada.

[10] Pant, Gautam., et al. (----). Search Engine-Crawler Symbiosis: Adapting to Community Interests. The University of Iowa, USA.

[11] Barker, Joe. (2003). What Makes a Search Engine Good? Available: http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SrchEngCriteria.pdf

[12] Bifet, Albert., et al. (2005). An Analysis of Factors Used in Search Engine Ranking. Technical University of Catalonia.

[13] Harth, Andreas, (----), Building a Semantic Web Search Engine: Challenges and Solutions, National University of Ireland, Galway.

[14] Types of Search Engines. Available: http://www.zeald.com/Resources/Promoting+&+Tuning+a+Successful+Web+Site/Types+of+Search+Engines.html

[15] Web Crawler Application Design. Available: http://searchenginecrawler.blogspot.com/2007/09/web-crawler-application-design.html

[16] The Anatomy Of An Automated Search Engine! Available: http://www.templatesfactory.net/articles/the-anatomy-of-an-automated-search-engine.html

[17] Finding a way to find your way on the Internet! Available: http://www.accessola2.com/superconference2005/thurs/docs/423/SearchDirDat.pdf

[18] Web Directories: a Selected List. Available: http://www.montgomerycollege.edu/library/webdirectories.pdf

[19] Santamarıa, Celina (2003), Automatic Association of Web Directories with Word Senses. Ciudad Universitaria, Madrid.

[20] System Analysis of Open Directory Project (ODP) http://dmoz.org. Available: http://www.ischool.utexas.edu/~i385tkms/blog/archives/ODP.pdf

[21] Calsavara, Alcides and Schmidt, Glauco (2004), Semantic Search Engines. Pontif´ıcia Universidade Cat´olica do Paran´, Brazil.

[22] Wen, Kunmei, et al., (2006). A Semantic Search Conceptual Model and Application in Security Access Control. Huazhong University of Science and Technology, China

[23] Rodrigo, L., et al (2005). A Semantic Search Engine for the International Relation Sector. Intelligent Software Components, S.A.

[24] Gruber, Thomas R. (1993), Toward Principles for the Design of Ontologies Used for Knowledge Sharing, Knowledge Systems Laboratory, Stanford University.

[25] Smith, Barry., Welty, Christopher (----), Ontology: Towards a New Synthesis, University at Buffalo & Vassar College, USA.

[26] Volker, Johanna., et al (2005), Automatic Evaluation of Ontologies (AEON), Institute AIFB, University of Karlsruhe

[27] Noy, Natalya F., McGuinness, Deborah L. (----),Ontology Development 101: A Guide to Creating

Authorized licensed use limited to: VIT-Amaravathi campus. Downloaded on September 25,2024 at 07:13:22 UTC from IEEE Xplore. Restrictions apply.

Your First Ontology, Stanford University, Stanford, CA.
[28] Application of Ontology Techniques to View-Based Semantic Search and Browsing
[29] Mudassar Ilyas, Qazi., et al. (2004). A Conceptual Architecture for Semantic Search Engine, University of Science and Technology, P. R. China.
[30] [31] Ontology Development 101: A Guide to Creating Your First Ontology. Available: http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
[32] Sanfilippo, Antonio., et al (2005) Automating Ontological Annotation with WordNet, Pacific Northwest National Laboratory, USA.
[33] Glossary: W. Available: http://www.archivemag.co.uk/glory/W.html
[34] Guha, R., et al. (----), Semantic Search, IBM Research Almaden, USA
[35] Semantic Search, Available: http://www.netlingo.com/lookup.cfm?term=semantic%20search
[36] Berkan, C., Riza, (2007). Semantic Search: An Antidote for Poor Relevancy. Available: http://www.readwriteweb.com/archives/semantic_search_antidote_for_poor_relevancy.php
[37] Celino, Irene, (----).Squiggle: a Semantic Search Engine for indexing and retrieval of multimedia content. Politecnico of Milano, Italy.
[38] A Semantic Search Engine for SRB Final Report (2007). Available: http://dart.edu.au/workpackages/si/si3-finalreport.pdf
[39] [40] Xerox unveils a new type of search engine, Available: http://www.bigmouthmedia.com/live/articles/xerox-unveils-a-new-type-of-search-engine.asp/3849/

Authorized licensed use limited to: VIT-Amaravathi campus. Downloaded on September 25,2024 at 07:13:22 UTC from IEEE Xplore. Restrictions apply.