EVALUATION OF TRADITIONAL AND SEMANTIC-BASED SEARCH ENGINES BASED ON THE SEMANTIC SEARCH PERFORMANCE

Harsh Tripathi, Akanksha Bindal

Delhi Technological University
Software Engineering Department
Delhi, India

ABSTRACT

This paper brings out the semantic search performance of traditional search engines and semantic-based search engines. Initially, four traditional search engines (Yahoo, Yandex, Dogpile and Google) and two semantic search engines (Bing, Kngine) are selected to compare their search performance on the basis of precision ratio and how they handle natural language queries. Twelve queries, from various topics were run on each search engine, the first thirty documents on each retrieval output was classified as being relevant or non-relevant. Afterwards, precision ratios were calculated for the first 30 document retrieved to evaluate performance of these search engines.Five natural language queries are then run on each of the aforementioned search engines to measure the relevancy of each retrieved document. These documents were classified as relevant or not relevant. The semantic search engines tend to handle these natural language queries with up to 70-90 percent precision while the traditional search engines fail to handle these natural language queries. Also, the authors inferred that Google is a traditional engine which is developing its semantic base at a rapid rate to be counted amongst the semantic engines in the near future.

Index Terms— Information Retrieval, Semantic Search, Engine,Keyword based Search Engine

I. INTRODUCTION

The search engine has become a primary need to explore the internet. A search engine is a web application system that is designed to search for information on the Web and retrieve the documents relevant to the users query. The most popular search engines are Google[1] , Yahoo[2] and Bing[3] with search volume ratios of 84.16 %, 7.61 %, 4.40 %, respectively[4]. The estimated size of the Web is at least 11.5 billion pages[5], but a much deeper(and larger) Web, estimated at over 3 trillion pages exists within databases whose contents the search engines do not index[6]. Although the research trend in the field of semantic search engines is increasing, publicly available search engines on the web are, generally, keyword-based search engines. Keyword-based search engines, such as Google and Yahoo, have a large user-base[7]. Therefore, evaluation of semantic search performance of popular keyword-based and semantic search engines is a valuable work intended to motivate researchers and search engine providers to improve current systems further. A perfect search engine model might be the one that always finds the precise document(s) on the web for the user. The result of a perfect search engine would, ideally, satisfy the expectations of its users, whenever a query is input. The inspiration, for this study, is to motivate researchers and search engine providers towards reaching this perfect search engine model. Traditional search engines are fundamentally different from semantic engines. This paper is organized as follows: Section 2 outlines the differences between a traditional and a semantic web search engine and delineates the reasons that have brought about a shift in the favor of semantic search all across the web. Section 3 a hands-on comparison (in terms of visual pictures) between the results returned by traditional and semantic engines for the same user query is depicted. Section 4 describes the methodology employed to evaluate search engines in terms of precision and normalized recall, Section 5 reports and discusses the experimental findings and the last section concludes the paper.

II. DIFFERENCE BETWEEN A TRADITIONAL AND SEMANTIC SEARCH ENGINE

II-A. Traditional Search Engines

Traditional Information Retrieval (IR) technology is based almost purely on the occurrence of words in documents. In this class of searches, the user provides the search engine a phrase or combination of words which s/he expects to find in the documents. There is no straightforward, reasonable interpretation of these words as denoting a concept.In such cases, the user is using the search engine as a navigation tool to navigate to a particular intended document[8].

- In the engine prompt, the user enters a keyword or sentence query.
- The engine does not understand polysemy and synonymy and the meaning of the terms.
- The system does not take into account stop words such as a, and, is, on, of, or, the, was, with, by, after, the.
- A traditional search engine is unable to handle long tail queries.

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 in 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 in a higher position in the search results page.

II-B. Semantic Search Engines

The availability of large amounts of structured, machine understandable information about a wide range of objects on the Semantic Web offers some opportunities for improving on traditional search. Semantic searches can overcome the 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[8].

- In the engine prompt of a semantic search engine, the user enters a question.
- The system understands polysemy and synonymy and knows the meaning of the terms. A semantic search engine takes into account stop words such as a, and, is, on, of, or, the, was, with, by, after, the.
- It is able to handle long tail queries.

It is designed to try and understand the context of the words that are used within the web page to try and match it more accurately to the user's search query.

II-C. Limitation of Keyword based Search Engine

Conventional Search Engines are very helpful in finding information on the internet and getting smarter with the passage of time, but they suffer from the fact that they do not know the meaning of the terms and expression used in the web pages and the relationship between them. Surveys indicate that almost 25% of Web searchers do not find adequate results in the first set of URLs returned, in part due to the daily sixty-terabyte increase in the size of the Web[9]. For our keyword based search engines, the amount of Web content out paces technological progress. In addition to their inability to keep pace with the growth of the Web, search engines rely too heavily on keyword-based string matching and word frequency and proximity techniques. As a result, queries are often overly sensitive to certain vocabulary used in the initial query string[10]. Search words often have multiple meanings or appear in multiple contexts, many of which are irrelevant to the Web searcher. Further, semantically similar pages that are desirable are often not retrieved, resulting in a set of results that is far from comprehensive. Some of the limitations of traditional search engines are:

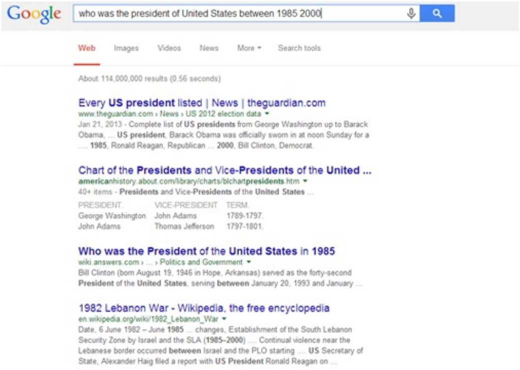
- Problem due to Polysemy words (one word having several meaning).e.g. Word Bank it can be a financial institution or river shore.
- Problem due to synonymy (several words having same meaning) e.g. For example, baby and infant are treated as synonyms in many thesauri, but Santa Baby has nothing to do with infant. Santa Babyis a song title, and the meaning of baby in this International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV131 entity is different than the usual meaning of infant[11].

III. SEMANTIC WEB SEARCH

The Semantic Web is an extension of the Web where information is represented in a machine process able way . While the information on the Web is mostly represented as HTML documents, RDF (Resource Description Framework) and OWL (Web Ontology Language)[12] are used for Semantic Web documents. In particular, the Semantic Web will contain resources corresponding not just to media objects (such as Web pages, images, audio clips, etc.) as the current Web does, but also to objects such as people, places,organizations and events. Further, the Semantic Web will contain not just a single kind of relation (the hyperlink) between resources, but many different kinds of relations between the different types of resources mentioned above[13].Semantic search engine like Hakia[14],Swoogle[15], DuckDuckGo[16] etc. are different from conventional search engines is that the semantic search engines are meaning based. Swoogle is a search engine for Semantic Web ontologys, documents, terms and data published on the Web. 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. Semantic search integrates the technologies of the Semantic Web and search engine to improve the search results gained by current search engines and evolves to the next generation of search engines built on the Semantic Web. 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. Semantic search integrates the technologies of the Semantic Web and search engine to improve the search results gained by current search engines and evolves to the next generation of search engines built on the Semantic Web.

For the comparison between Traditional Search Engines vs. Semantic Search Engine, we use two search engines Kngine, as a semantic search engine and Google, as a traditional search engine. The semantically relevant phrase who was the president of the United States between 1985-2000 is used as search query. See the result in Figure 1(a) and figure 1(b).

From the figures 1(a) and (b) it is clear that Kngine returns nicely visual results that directly answer the user query. It does so because it understands the semantics of the phrases you typed. Google, by contrast, returns thousands of



who was the president of United States between 1985 2000

About 114,000,000 results (0.56 seconds)

Every **US president** listed | News | theguardian.com
www.theguardian.com › News › US 2012 election data
Jan 21, 2013 - Complete list of **US presidents** from George Washington up to Barack Obama, ... **US president**, Barack Obama was officially sworn in at noon Sunday for a .... **1985**, Ronald Reagan, Republican ... **2000**, Bill Clinton, Democrat.

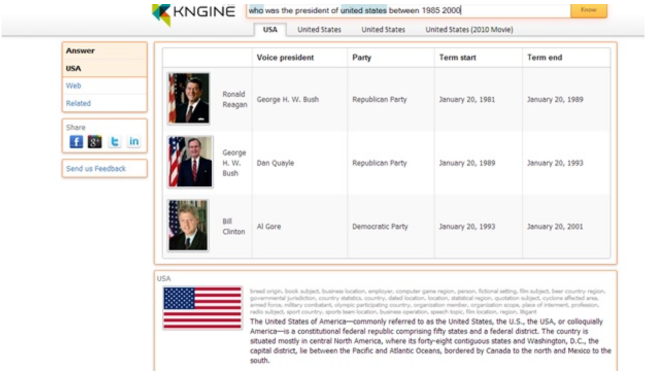
Chart of the **Presidents** and Vice-**Presidents** of the **United** ...
americanhistory.about.com/library/charts/blchart**presidents**.htm
40+ items - **Presidents** and Vice-**Presidents** of the **United States** ...

PRESIDENT.	VICE-PRESIDENT	TERM.
George Washington	John Adams	1789-1797.
John Adams	Thomas Jefferson	1797-1801.

Who was the President of the **United States** in **1985**
wiki.answers.com › ... › Politics and Government
Bill Clinton (born August 19, 1946 in Hope, Arkansas) served as the forty-second **President** of the **United States**, serving **between** January 20, 1993 and January ...

1982 Lebanon War - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/1982_Lebanon_War
Date, 6 June 1982 – June **1985** ... changes, Establishment of the South Lebanon Security Zone by Israel and the SLA (**1985–2000**) .... Continual violence near the Lebanese border occurred **between** Israel and the PLO starting .... **US** Secretary of State, Alexander Haig filed a report with **US President** Ronald Reagan on ...

(a)Search result by Google search engine



who was the president of united states between 1985 2000

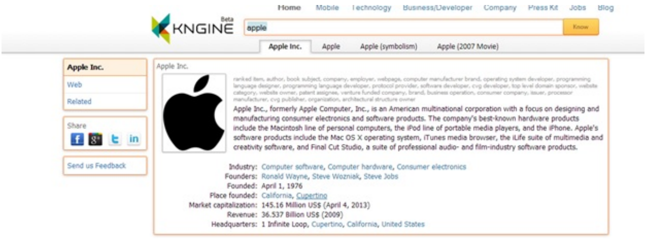
	Voice president	Party	Term start	Term end
Ronald Reagan	George H. W. Bush	Republican Party	January 20, 1981	January 20, 1989
George H. W. Bush	Dan Quayle	Republican Party	January 20, 1989	January 20, 1993
Bill Clinton	Al Gore	Democratic Party	January 20, 1993	January 20, 2001

USA

breed origin, book subject, business location, employer, computer game region, person, fictional setting, film subject, beer country region, governmental jurisdiction, country statistics, country, dated location, location, statistical region, quotation subject, cyclone affected area, armed force, military combatant, olympic participating country, organization member, organization scope, place of interment, profession, radio subject, sport country, sports team location, business operation, speech topic, film location, region, litigant

The United States of America—commonly referred to as the United States, the U.S., the USA, or colloquially America—is a constitutional federal republic comprising fifty states and a federal district. The country is situated mostly in central North America, where its forty-eight contiguous states and Washington, D.C., the capital district, lie between the Pacific and Atlantic Oceans, bordered by Canada to the north and Mexico to the south.

(b)Search result by Kngine (semantic) search engine



apple

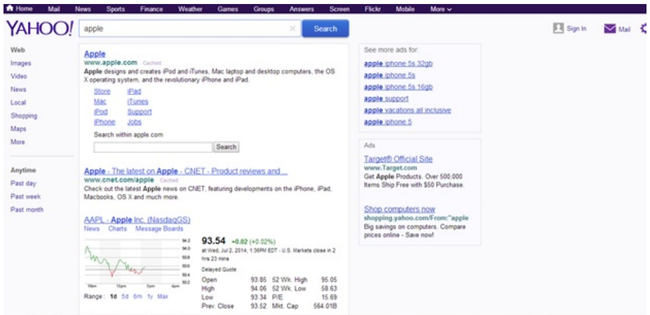
Apple Inc.

ranked item, author, book subject, company, employer, webpage, computer manufacturer brand, operating system developer, programming language designer, programming language developer, protocol provider, software developer, cvg developer, top level domain sponsor, website category, website owner, patent assignee, venture funded company, brand, business operation, consumer company, issuer, processor manufacturer, cvg publisher, organization, architectural structure owner

Apple Inc., formerly Apple Computer, Inc., is an American multinational corporation with a focus on designing and manufacturing consumer electronics and software products. The company's best-known hardware products include the Macintosh line of personal computers, the iPod line of portable media players, and the iPhone. Apple's software products include the Mac OS X operating system, iTunes media browser, the iLife suite of multimedia and creativity software, and Final Cut Studio, a suite of professional audio- and film-industry software products.

Industry: Computer software, Computer hardware, Consumer electronics
Founders: Ronald Wayne, Steve Wozniak, Steve Jobs
Founded: April 1, 1976
Place founded: California, Cupertino
Market capitalization: 145.16 Million US$ (April 4, 2013)
Revenue: 36.537 Billion US$ (2009)
Headquarters: 1 Infinite Loop, Cupertino, California, United States

(c)Search Result of keyword apple by Kngine Search Engine



apple

Apple
www.apple.com Cached
Apple designs and creates iPod and iTunes, Mac laptop and desktop computers, the OS X operating system, and the revolutionary iPhone and iPad.
Store iPad
Mac iTunes
iPod Support
iPhone Jobs
Search within apple.com
Search

Apple - The latest on **Apple** - CNET - Product reviews and ...
www.cnet.com/apple Cached
Check out the latest **Apple** news on CNET, featuring developments on the iPhone, iPad, Macbooks, OS X and much more.

AAPL - **Apple** Inc. (NasdaqGS)

93.54 +0.02 (+0.02%)
at Wed, Jul 2, 2014, 1:36PM EDT - U.S. Markets close in 2 hrs 23 mins
Delayed Quote

Open	93.85	52 Wk. High	95.05
High	94.06	52 Wk. Low	58.63
Low	93.34	P/E	15.69
Prev. Close	93.52	Mkt. Cap	564.01B

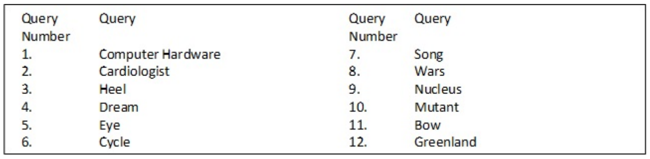
See more ads for:
apple iphone 5s 32gb
apple iphone 5s
apple iphone 5s 16gb
apple support
apple vacations all inclusive
apple iphone 5

Ads
Target® Official Site
www.Target.com
Get **Apple** Products. Over 500,000 Items Ship Free with $50 Purchase.

Shop computers now
shopping.yahoo.com/From:"apple
Big savings on computers. Compare prices online - Save now!

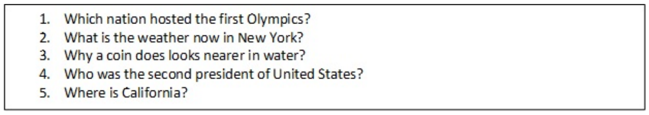
(d)search result of keyword apple by Yahoo Search Engine

Fig. 1. Search Engine Results



Query Number	Query	Query Number	Query
1.	Computer Hardware	7.	Song
2.	Cardiologist	8.	Wars
3.	Heel	9.	Nucleus
4.	Dream	10.	Mutant
5.	Eye	11.	Bow
6.	Cycle	12.	Greenland

(a)Queries Used



1. Which nation hosted the first Olympics?
2. What is the weather now in New York?
3. Why a coin does looks nearer in water?
4. Who was the second president of United States?
5. Where is California?

(b)Natural Language queries

Fig. 2. Methodology

results for these queries, but you wont easily find any results responding directly to these queries. Articles related to the Lebanon War are also listed as relevant as Google and other search engines rely mostly on understanding keywords, and theyve had difficulty moving beyond them.

From the figures 1(c) and (d) it is clear that Yahoo is not able to handle Polysemy words while Kngine which is a Semantic Search Engine provide results in all possible meaning to the word Apple. Apple can be a software giant(Apple Inc), fruit, symbolism(as the forbidden fruit in many religions and traditions), the 2007 Swedish Movie, etc.

IV. METHODOLOGY

Four traditional search engines, namely Yahoo, Google,Yandex and Dogpile, and two semantic search engines, namely Kngine and Bing, have been researched upon. Thereafter, twelve keyword queries that consists of various topics and contain one or more keywords, given in Figure 2(a), were searched upon the engines. After each run of the query, the first 30 documents retrieved were evaluated using binary human relevance judgement and with this, every document was classified as relevant or non-relevant. Precision ratios of keyword based search engines were calculated for first 30 documents retrieved for each pair of query and search engine.

Precision and recall are calculated as:

1) $$precision = \frac{relevantdocumentsretreived}{totaldocumentsretreived}$$

2) $$recall = \frac{relevantdocumentsretreived}{totalrelevantdocuments}$$

V. EXPERIMENTAL RESULTS

The number of relevant documents retrieved by each search engine for the first thirty documents is shown in the Figure 3(a) along with the precision rate for all the engines. In the research carried out by the authors, it was found that the precision rate of Bing was highest (0.744) amongst all the engines used and the precision rate of Yandex was the lowest (0.4722). Also, Google, although being a traditional search engine gave close to semantic results

(0.675), which implies that it is moving towards semantic search engine category. Figure 3(c) represents the Precision ratio for the first 30 documents. While, Figure 3(d) represents graphical representation for the twelve queries (number 1 to 12). Thereafter, natural language queries were run on the various search engines and the following results were obtained, where Y represents relevant result to the query and N represents irrelevant result to the query, represented in Figure 3(b).

VI. CONCLUSION

The research shows that the growth in semantic search technology and the use of natural language processing of the search engines has greatly impacted information technology and changed the search engines information retrieval techniques. In this paper, detailed research has been carried out on the semantic search performance of traditional search engines and semantic-based search engines. Bing retrieved the maximum relevant documents followed by Knigne. Whereas, Yandex gave the most irrevlant results leading to a precision of less than 0.45. Dogpile failed to retrieve any relevant document for all the language queries, while Yahoo and Yandex, both failed in three natural language searches out of five. Bing and Kngine were able to retrieve a relevant answer for all the queries while Google failed to retrieve a relevant answer for only one natural language query. The high precision rate(0.675) and exceptional handling of natural language queries by Google indicate that despite being a traditional search engine, Google is moving towards the bracket of Semantic Search engines based on its advancement in searching technology.

VII. REFERENCES

[1] "Google." [Online]. Available: http://www.google.com

[2] "Yahoo." [Online]. Available: http://www.yahoo.com

[3] "Bing." [Online]. Available: http://www.bing.com

[4] "Kngine." [Online]. Available: http://www.kngine.com

[5] J. Singh and A. Sharan, "A comparative study between keyword and semantic based search engines."

[6] A. Gulli and A. Signorini, "The indexable web is more than 11.5 billion pages," in *Special interest tracks and posters of the 14th international conference on World Wide Web*. ACM, 2005, pp. 902–903.

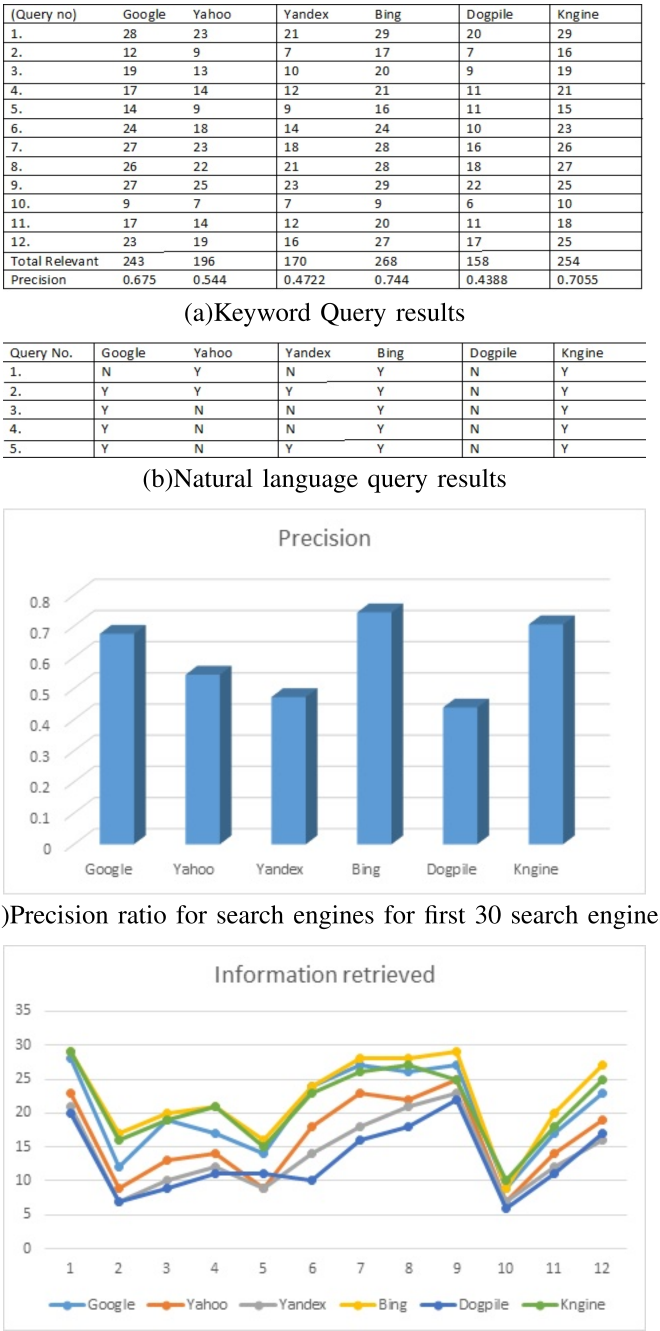
[7] D. Tumer, M. A. Shah, and Y. Bitirim, "An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia," in *Internet Monitoring and Protection, 2009. ICIMP'09. Fourth International Conference on*. IEEE, 2009, pp. 51–55.

[8] "Hitwise." [Online]. Available: http://www.hitwise.com/datacenter

[9] W. Roush, "Search beyond google," *TECHNOLOGY REVIEW-MANCHESTER NH-*, vol. 107, no. 2, pp. 34–45, 2004.

[10] G. Antoniou and F. Harmelen, "A semantic web primer, chapter 1," 2004.

[11] X. Wei, F. Peng, H. Tseng, Y. Lu, X. Wang, and B. Dumoulin, "Search with synonyms: problems and



(Query no)	Google	Yahoo	Yandex	Bing	Dogpile	Kngine
1.	28	23	21	29	20	29
2.	12	9	7	17	7	16
3.	19	13	10	20	9	19
4.	17	14	12	21	11	21
5.	14	9	9	16	11	15
6.	24	18	14	24	10	23
7.	27	23	18	28	16	26
8.	26	22	21	28	18	27
9.	27	25	23	29	22	25
10.	9	7	7	9	6	10
11.	17	14	12	20	11	18
12.	23	19	16	27	17	25
Total Relevant	243	196	170	268	158	254
Precision	0.675	0.544	0.4722	0.744	0.4388	0.7055

(a)Keyword Query results

Query No.	Google	Yahoo	Yandex	Bing	Dogpile	Kngine
1.	N	Y	N	Y	N	Y
2.	Y	Y	Y	Y	N	Y
3.	Y	N	N	Y	N	Y
4.	Y	N	N	Y	N	Y
5.	Y	N	Y	Y	N	Y

(b)Natural language query results

(c)Precision ratio for search engines for first 30 search engines

(d)Visual Representation of Search engines for the twelve queries

Fig. 3. Experimental Results

solutions," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010, pp. 1318–1326.

[12] J. M. Kassim and M. Rahmany, "Introduction to semantic search engine," in *Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on*, vol. 2. IEEE, 2009, pp. 380–386.

[13] I. Celino, E. Valle, D. Cerzza, and A. Turati, "Squiggle: a semantic search engine for indexing and retrieval of multimedia content," *Proceedings of SAMT*, vol. 2006, pp. 20–34, 2006.

[14] "Hakia." [Online]. Available: http://www.hakia.com

[15] "Swoogle." [Online]. Available: http://www.swoogle.umbc.edu

[16] "Duckduckgo." [Online]. Available: http://www.duckduckgo.com