Journal of Advanced Computing and Communication Technologies ***(ISSN:* 2347 - 2804*)***
Volume No 4 Issue No.2, April 2016

Comparative Analysis of Semantic and Syntactic Based Search Engines

By

Sanjib Kumar Sahu, Nitin Kumar, Saurabh Allawadhi, Japneet Singh
Department of Computer Science, Utkal University, Bhubaneswar, Odisha, India
USICT, Guru Gobind Singh Indraprastha University, New Delhi, India
USICT, Guru Gobind Singh Indraprastha University, New Delhi, India
USICT, Guru Gobind Singh Indraprastha University, New Delhi, India
sahu_sanjib@rediffmail.com, nitinrajput053@gmail.com, saurabhallawadhi029@gmail.com, japneetheyer@yahoo.in

ABSTRACT
Semantic based search engines are better than syntactic or key word based search engines as they have the ability to understand the meaning of the searched term and provide more specific data. We studied semantic (Hakia, Kosmix, Cognition, Lexxe, Swoogle) and syntactic (Google, Yahoo, Ask) based search engines and compared their search performance and analyzed them using specific queries. We also focused on some specific algorithms i.e. Page Rank algorithm, Weighted Page rank algorithm, Tag search algorithm.

Keywords
Semantic search engines, Keyword search engines, Information retrieval, and Page rank

1. INTRODUCTION

1.1 Brief Introduction
A search engine can be described as a system that is developed mainly for searching information on W3(World Wide Web). The main need of the search engine is to explore the data present on the internet. If the concept of search engine would not have been there then it would have become very difficult for the users to search data on so many websites on the internet. There are many search engines used like Google, Yahoo, Ask, Bing etc for finding the useful information. As the internet has grown many search engines have also been developed to give variety to users for searching the information but only some search engines provide the specific information related to the search of the user. Automated clustering of data is done to manage the large number of web related documents and also the popularity has increased by organizing them into domain dependency directories. The web has evolved a lot in the past years in the form of generations in which web1.0 was the first generation starting from 1990 and ending in 2000, web 2.0 was the second generation starting from 2000 and ended in 2007 to web 3.0 which started in 2007 and is also currently being used.

1.2 Web Versions: an Overview

1.2.1 Web 1.0 Version

It can also be described as the first satge of the world wide web.It is basically a collection of web pages connected to each other by the use of hyperlinks. The correct definition of the first version of web is still a source of a talkpoint but it is referred as a collection of static web sites which did not contained much interactive content in them. The first version also includes static information in it.

Fig 1: Graphical View of Web 1.0

1.2.2 Web 2.0 Version

The second version of web purely described the W3(World Wide Web) sites which mainly had the user generated content, their usability and interoperability in the web sites . There were two main personalities who were responsible for popularizing the second version of web at the O'Reilly Media Web 2.0 Conference in 2004 named as Tim O'Reilly and Dale Dougherty. But in 1999 it was mainly coined by Darcy DiNucci.

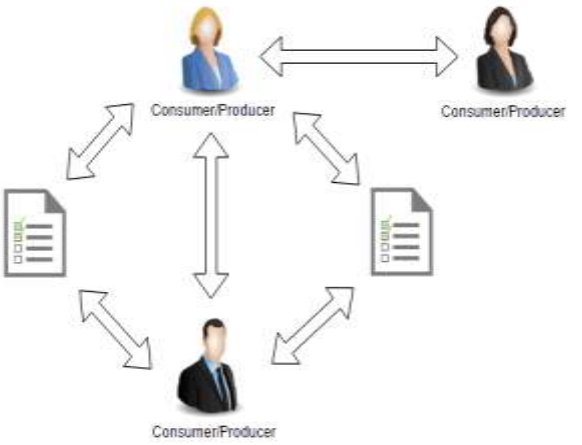

Fig 2: Graphical View of Web 2.0

1.2.3 Web 3.0 Version

It can be defined as an extension to the second version of web and the evolution of the web in the world of internet. Tim O'Reilly famously gave this very popular view about the definition of the Third version of web. But, a person named as Nova Spivack described this version as the source of connecting the applications, data and the people using the internet. Also, some people define the third version of web as one of the fast developing technology and much more to come in it.

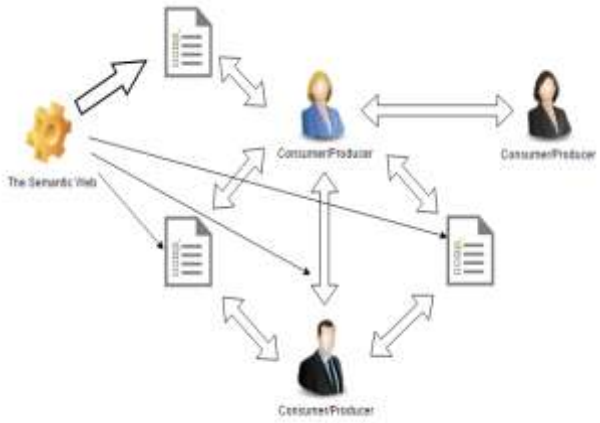

Fig 3: Graphical View of Web 3.0

The most common activity which is being done by large number of users on the internet is searching and it is being carried out by large number of audience working on the web. Research has been carried out in past years for making the use of search engines easy. Now days, the search engine is made to work on the semantic occurrences. It basically improves the searching quality on the basis of understanding the need of the user and also to understand the meaning of the tokens as they appear on the web or within a closed system which helps in generating more specific and useful results for the user.

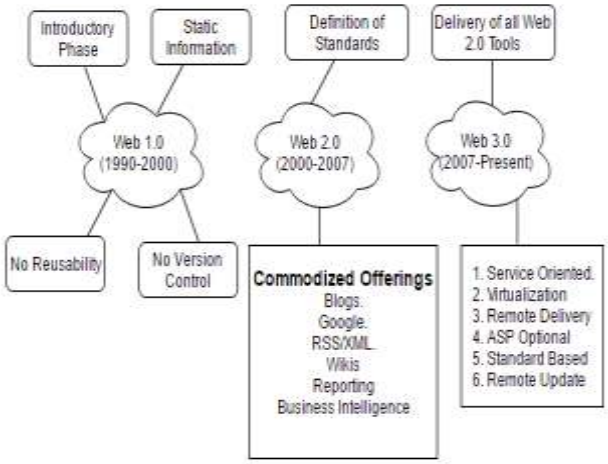

Fig 4:
Relationship between Web 1.0, Web 2.0 and Web 3.0

The roadmap to the remaining part of the paper is as follows. In Section 2 we discussed the different semantic search engines and their brief explanation. Section 3 proposes the different syntactic search engines and related studies. Section 4 we are discussing about comparison between the semantic search engine and syntactic based search engines. Section 5 presents the information about the proposed system. Finally, finally we concluded the paper and defined the future scope in Section 6.

2. SEMANTIC SEARCH ENGINES

A Semantic Search Engine consists of important features:

- Development based on Ontology.
- Crawler based on Ontology.
- Expert System based on Ontology
- Semantic Search Performance.
- Query Processing and making.

Ontology is critical concept that must be studies as it helps in understanding of domain in following forms:

(1) Structure of Domain.

(2) Reusability of Domain.

(3) Making Domain Assumptions.

(4) Analyze Domain knowledge.

(5) Differentiate between domain and operational knowledge.

Semantic based search engines are typically uses a web crawler that make it more helpful as it automatically scans over the web pages just like a spider. So this phenomenon is known as web crawling or spidering.

Now a ontology translator converts web pages scanned by crawler into textual documents with the help of a mapper so that their indexing and storing becomes efficient. A typical

relational database tables are utilized here for storing these tables. [7][8]

2.1 Hakia

1. Founded by Riza Berkan.
2. Founded on 2004.
3. It used a different method for indexing that is called QDEXing technology.
4. It's headquarter is in New York city.
5. Since April 2004 hakia.com has been offline.[14].

2.2 Kosmix

1. Founded on 2005.
2. Founded by Anand Rajaraman, Venky Harinarayan.
3. Kosmix now known as WalmartLabs. It was taken over by Walmart in April 2011 and made a research department called WalmartLabs.
4. American firm situated in Mountain View, California.5. Website mostly earns from advertisements that is linked to categorization.
5. Website allow users to view web pages as topic pages that contains videos, photos, news, commentaries and other links.

2.3 Cognition

1. Cognition semantic searching is based on algorithms that help in reading the documents.
2. These are to eliminate different
   a) Sense.
   b) Forms.
3. Concepts of a word.
4. Searching gives five basic approaches
   a) Typical English
   b) Fuzzy logic
   c) Boolean (yes or no)
   d) Phrase
   e) Pattern mapping

2.4 Lexxe

1. Founded by Dr. Hong Liang Qiao.
2. Founded on 2005.
3. Uses natural language processing in semantic searching.
4. Company situated Sydney, Australia.
5. Alpha version
   a) Launched In 2005.
   b) Include keyword searching and searching in form of questions in simple English
   c) Stopped for further development in 2010.
6. Beta version
   a) Launched in 2011.
   b) Searching now based on typical keyword searching and keyword that were of some special meaning or we can say keyword are based on specific concept hence the name Lexxe semantic keys searching was given.
   c) This version closed on June 17, 2015 for further analysis.

2.5 Swoogle

1. Founded by Li Ding and Tim Finin.
2. It mainly detects RDF documents HTML pages that are linked with RDF documents.
3. It stores these pages into database by indexing them and store them in the form of metadata.
4. It uses Page Rank algorithm made by Google for ranking purposes and it also adds the feature of semantic web documents to make Swoogle semantic based search engine.

3. SYNTACTIC BASED SEARCH ENGINE

Syntactic based search is used to search current information mainly from the HTML content. Also the web pages are not read in a systematic predefined order.

The syntactic based search does not follow special tagging techniques like page rank algorithm or tag search algorithm which makes it difficult for computer program to search them and arrange them according to specific order.

They are also less reliable than semantic based engines as they are not able to understand the meaning of the term searched by the user and it provides the results in unspecific order.

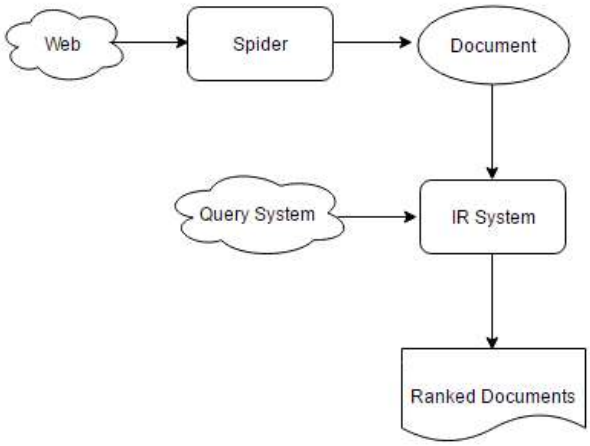

Fig 5: Information retrieval process in Syntactic based search engine

Various Syntactic based search engines are as follows:

3.1 Google

1. Founded on September 4, 1998.
2. Founded by Larry Page and Sergey Brin.
3. Google is an American multinational technology company dealing in Internet-related services and products.

4. It also deals with online advertising technologies that is Ad Words that puts advertisement near the searched answers, cloud computing, searching and software products.

3.2 Yahoo

1. Founded on January 1994.
2. Founded by Jerry Yang, David Filo.
3. Yahoo Inc. (styled as Yahoo!) is an American multinational technology company.
4. It is best known for search engine and their web related services.
5. It mainly includes mail, news, advertisement services etc.

3.3 Ask

1. Founded on June 1996.
2. Founded by Garrett Gruener and David Warthen.
3. Ask.com (previous known as AskJeeves.com) is a basic question-answer website.
4. It was first implemented by Warthen, Chevsky, and Justin Grant by taking Gary Chevsky's core design.
5. But due to increasing competition in late 2010 from other search engines they outsourced their work as web search technology and just focused on question answer concept.

4. COMPARISONS BETWEEN SEMANTIC BASED SEARCH ENGINES AND SYNTACTIC BASED SEARCH ENGINES

Table 1: Comparison between the semantic and syntactic based information retrieval system

Information Retrieval System	Semantic Information Retrieval System	Syntactic Information Retrieval System
Keywords	Some value, min, exactly, only, max	AND, OR and NOT
Symbol	∃ , ∋, ≥ , =, ≤ , ∀	+, -, ()
Phrase	[], “”	“”
Case Sensitive	YES	NO
Wildcards	*, ?, $	(*)
Prefixes	Length, Max & Min Length, Fraction & Total Digits.	File type,

Table 2: Comparative study on ranking techniques

Algorithm	Concept	Ranking Basis	Advantages
Page Rank Algorithm	In links are used	Importance of pages	No stress on personalization
Weighted Page Rank Algorithm	In links & out links are used	Popularity of pages	No stress on personalization
Tag Search Algorithm	Social Hints	Social Tagging	Cannot depend on social hints only.

5. PROPOSED SYSTEM

We have proposed a framework such that semantic based Search engines retrieve more relevant information. We propose an engine 'X' that provide relevant and reliable searched web pages to the user. Firstly the query is made to translator that translates it to natural language. It is then fed to question-answer system which helps in decreasing the number of pages to be displayed to the user. This helps in reducing the number of comparisons that has to be made into crawler. Crawler is basic system that consists of data repositories, data translator and database. Relational database and data mining techniques is used to fasten up the process. Repositories are utilized to compare the query to past data so that proper web pages can be displayed. To segregate the web pages that are to be displayed is done using question-answer system and the crawler. We utilized Q&A system as it is very easy to implement. The crawler system mainly performs two functions that are (1) capture links related to searched term and (2) store them into repositories in textual format and database so that comparisons can be made efficiently and precisely.

Our proposed framework describes important facts and methods that are helpful in finding results and it uses general concepts so that searching is fast and transparent.

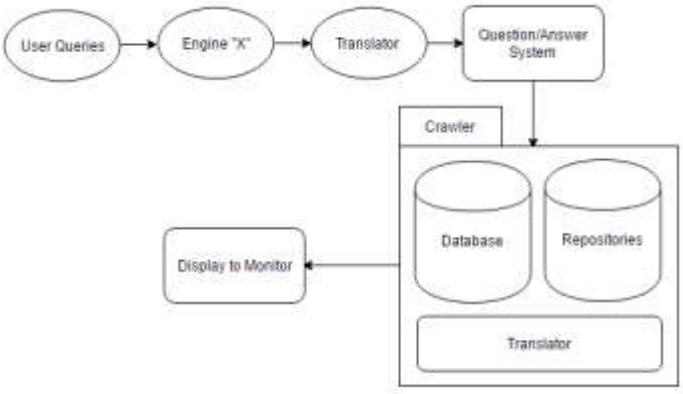

Fig 6: Block Diagram of the Proposed System

From Figure 7-11 it presents the search results for the syntactic based search engines –Google, Yahoo, Ask and semantic based search engines – Cluuz and Swoogle. It also shows the number of links for the search engines.

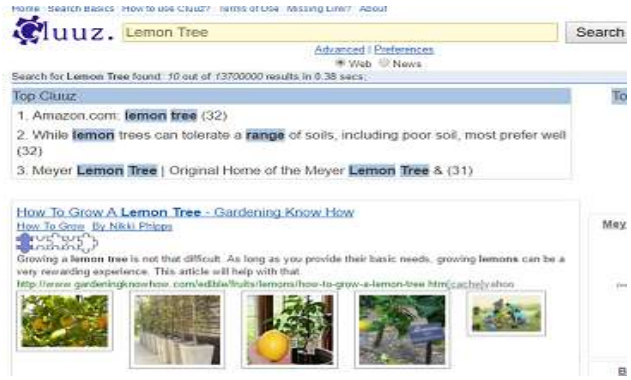

Fig 7: Cluuz based search results

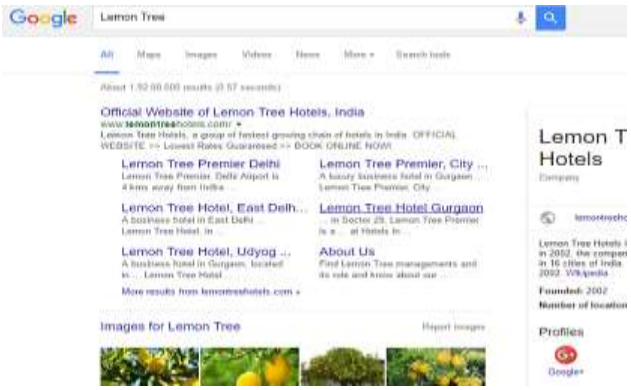

Fig 8: Google Based Search results

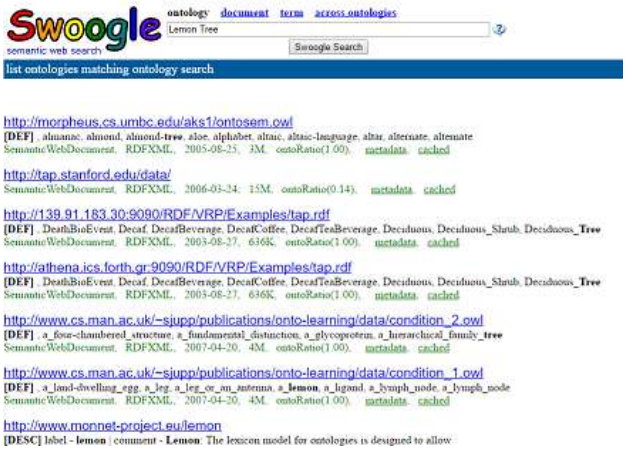

Fig 9: Swoogle based search results

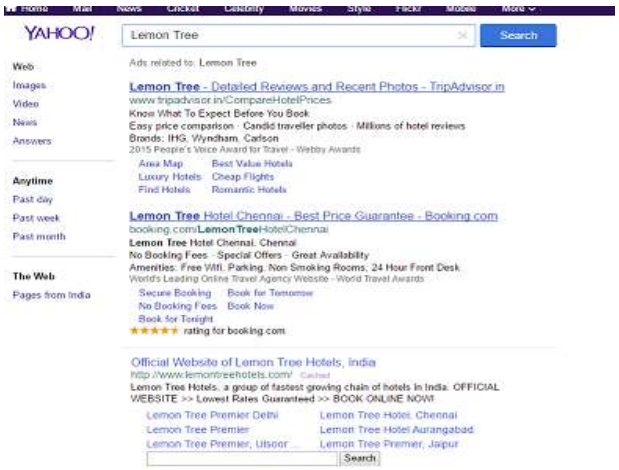

Fig 10: Yahoo based search results

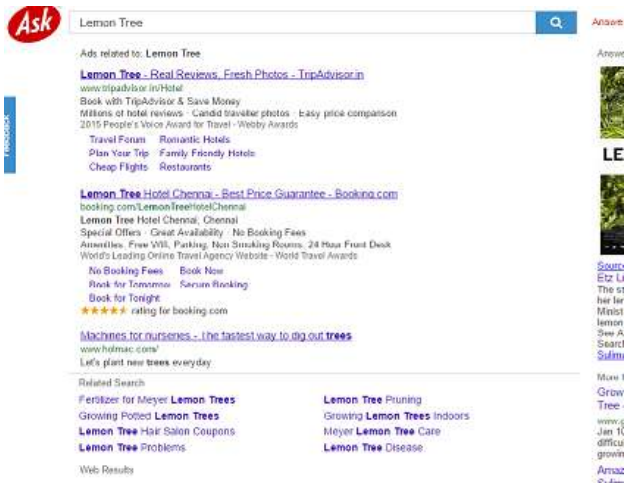

Fig 11: Ask based Search results

6. CONCLUSION AND FUTURE SCOPE

Information Retrieval is a critical concept now a day as the data is increasing as time passes on and the user wants data fastly, correctly, reliably, efficiently and precisely.

Data is handled using various clouds but there had always been an urge to fetch the data as early as possible as time is money.

So we studied these various techniques available and the various existing search engines either semantic or syntactic and tried to propose a real life search engine that will solve problems like searching optimized results that are not well tagged in a specific order.

From studying various semantic based search engines and syntactic based search engines, we concluded that the each search engine that we studied are best in their specific respect just like Google is best for fast searching by using spider algorithm.

We proposed a search engine in which we incorporated the best features so that searching becomes precise, better and efficient.

7. 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] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, pp. 29–37, 2001.

[3] www.toptensemantic.com (top ten semantic web sites)

[4] Wang wei, Payam M. Barnaghi and Andrzej Bargiela , Semantic enhanced information search and retrieval, Advanced Language Processing and Web Information Technology, 2007. ALPIT 2007. Sixth International Conference on 2007.

[5] http://www.searchenginejournal.com/semantic-search-engines/9832/

[6] http://www.w3.org/wiki/Search_engines

[7] Ontology Development 101: A Guide to Creating Your First Ontology.

[8] Sanfilippo, Antonio., et al (2005) Automating Ontological Annotation with WordNet, Pacific Northwest National Laboratory, USA.

[9] Application of Ontology Techniques to View-Based Semantic Search and Browsing

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

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

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

[13] Wang wei, Payam M. Barnaghi and Andrzej Bargiela , "Semantic enhanced information search and retrieval", Advanced Language Processing and Web Information Technology, 2007. ALPIT 2007. Sixth International Conference on 2007.

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

[15] S. S. Al-rawi, A. T. Sadiq, and S. A. Hamad, "Design and Evaluation of Semantic Guided Search Engine." International Journal of Web Engineerung, 1(3), pp. 15–23, 2012.

[16] G. Sudeepthi, G. Anuradha, P. M. Surendra, and P. Babu, "A Survey on Semantic Web Search Engine," vol. 9, no. 2, pp. 241– 245, 2012.

[17] G. Madhu, a Govardhan, and T. K. V. Rajinikanth, "Intelligent Semantic Web Search Engines: A Brief Survey," International journal of Web & Semantic Technology, vol. 2, no. 1, pp. 34–42, Jan. 2011.

[18] P. Mika, "Microsearch_: An Interface for Semantic Search,"2008