



## Transition in Web Search to a Semantic Perspective

Aatif Ahmad Khan<sup>1</sup> & Sanjay Kumar Malik<sup>2</sup>

<sup>1</sup>Research Scholar, University School of Information, Communication & Technology

<sup>2</sup>Associate Professor, University School of Information, Communication & Technology  
Guru Gobind Singh Indraprastha University, New Delhi, India

**Abstract:** Incorporation of semantic elements into traditional web search is improving its efficiency in terms of search results. Result set of search query that used to have long list of syntactically matching web documents is being augmented with contextually relevant and highly precise information. Further, presentation of semantic knowledge using schemes such as recommendation sidebars, auto-complete suggestions etc. has powered users to explore conceptually linked data for a specified search query. The need is to explore the transition from keyword-based search to semantics-based search along with incorporation of semantic technologies into web search and applications.

In this paper, first, transition from keyword-based search to meaning-based search on the web with augmentation of semantic web technologies is explored along with various limitations and issues of keyword-based search. Secondly, traditional approaches such as Page Rank algorithm, has been illustrated. Then, applications of semantic web technologies and NLP techniques for advancements in web search are explored, also considering the availability of semantic search engines. Thirdly, various semantic result presentation schemes including recommendations and Wikipedia Infoboxes are reviewed.

**Keywords:** semantic search, information retrieval, semantic web technology, semantic knowledge presentation, recommendation sidebar, Wikipedia Infobox

### I. Introduction

With availability of the World Wide Web to mass public, the scope of web search has increased tremendously. Internet users, both humans and automated bots, use web search for fulfilling their specific information needs. Web search providers have gathered huge popularity for their capability to collect information from diverse sources and returning information ranked in order of relevance in response to a search query. Such service providers including Google, Bing etc. also maintain caches for improving query response times.

But, in this age of Big Data and ever increasing size of web resources, efficiency of web search is being affected in terms of relevancy of results. Sometimes, irrelevant results are also included as top results due to lack of context understanding on machine part. This also accounts due to usage of search terms which are ambiguous in nature (may exhibit different meanings when used in different contexts). E.g. keyword “duck” may mean a bird but when used in context of cricket, it means batsman got out without scoring any run. Hence the query “number of ducks in India” may require additional context information such as cricket or an actual bird. Further, results are affected by synonyms present in web documents. These documents will be marked as irrelevant if technique used is keyword matching alone, but these documents may be highly relevant conceptually. Above issues imply that keyword only searches should not be utilized as the only criteria for web search. Although this approach may suffice for navigational search [7], i.e. type of search in which user only expects these keywords to be found in web document. But, apart from navigational search, we need more sophisticated approaches than keyword only search.

With the vision of Semantic Web, its technologies, and progress attained in these directions, machines are being made intelligent enough to understand underlying meaning of sentences. This direction is further benefited by Natural Language Processing (NLP) techniques. Using advanced sentence analysis techniques, correct sense can be assigned to ambiguous words. Various semantic web technologies such as ontology, RDF etc. provides the basic building blocks of incorporating semantics into search techniques. Ontology consists of concept hierarchies and interrelationships among concepts. These are often used in deriving semantic relatedness among terms. Resource Description Framework (RDF) adds structure to data using semantic metadata. RDF/XML adds both syntax and semantics to web data. Also, NLP techniques when used along with semantic web technologies may further increase the efficiency of search as these can be used to overcome linguistic issues such as synonymy, homonymy, hyponymy etc.

At present, this incorporation of semantics into web search can be readily seen in logical results returned by popular search engines. Relevant recommendations can also be seen appended to list of web pages. Also some projects which use semantics into result presentation are there such as Wikipedia Infoboxes.

#### A. Issues/Limitations with keyword based search

Some key issues involved with keyword-based search are described in Table 1:

**Table 1: Key issues with keyword-based search**

| Issue                                        | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|----------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Massive Size of Web                          | Massive size of web having documents containing keywords all around, hence even inverted indices will be containing enormously large number of documents for corresponding keywords. Also for such an enormous collection, feasibility of using plain indices is impractical.                                                                                                                                                                                                                                                                                                   |
| Efficiency Constraints (Query Response Time) | Efficiency constraints due to large number of matching documents (Search Engine can't search keywords in all documents else query response time will increase tremendously). To improve response times, search engines often maintain caches of response webpages for particular queries, but these caches needs to be updated frequently.                                                                                                                                                                                                                                      |
| Ambiguous Keywords                           | There is often multi-meaning mapping of keywords which are used in the search queries. For example the keyword "mean" may be treated differently in different contexts: <ul style="list-style-type: none"> <li>• What do you <i>mean</i>?</li> <li>• You are so <i>mean</i>!</li> <li>• The <i>mean</i> of first 5 natural numbers in 3.</li> </ul> For such queries containing ambiguous terms, search engines use Word Sense Disambiguation (WSD) techniques to get the context of query. Often, some more information from user is required to further filter the responses. |
| Synonyms                                     | Keyword-based search treats synonyms as totally different keywords. This limitation may flag a highly relevant document as irrelevant one due to keyword mismatch.                                                                                                                                                                                                                                                                                                                                                                                                              |
| Linguistic barriers                          | Lack of machine intelligence for hyponymy, hypernymy and homonymy E.g. "Red" is a "Color" but search query with keyword "Color" may treat documents containing "Red" as irrelevant.                                                                                                                                                                                                                                                                                                                                                                                             |
| Filtering Content                            | Inability to narrow down (filter) search results, due to inability to understand context.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |

Also, it can't be just assumed that text-based search capabilities will be sufficient for searching over the web because comparing the web with the text-based databases shows following differences [9]:

- Web is a knowledge base for countably many areas, searching it will require advanced text understanding as vocabulary corresponding to an area may overlap with another area, but containing huge difference in meanings.
- Web is somewhat structured (uses HTML etc.) and interlinked, so understanding of some content may also be derived using other contents that point to it or are pointed by it.
- Number of documents grow rapidly on the web and their contents also change frequently (so that an inverted page index maintained by search engine may not be relevant over a period of time if specific keywords contained in them gets updated over time).
- Web resources are highly unstructured.
- Due to enormous size of web, simple text queries may result in massively large number of matching documents.
- User wishes to find specific information quickly, hence ranking is a key factor for search over web.
- Documents on web are available in many formats (HTML, PDF, JPG etc.) and may contain different media types as their contents (text, images, audio, video etc.).

This paper explores and revisits the advances in the direction of web search from a Semantic Information Retrieval (IR) perspective. It includes transition from keyword-based search to meaning-based search on the web with augmentation of semantic web technologies and various limitations and issues of keyword-based search are also explored. In section 2, Page Rank, a traditional approach for IR over web has been illustrated. In section 3, an overview of semantic search along with applications of various semantic web technologies and NLP techniques for semantic search is explored. A brief scope and applicability of semantic search engines is also discussed in this section. In Section 4, some practical implications of incorporation of semantic web technologies into IR over web search such as recommendation sidebars, Wikipedia Infoboxes etc. are presented. Finally, a discussion across key insights gathered in this paper is presented in Section 5. It is followed by Future Scope in this direction.

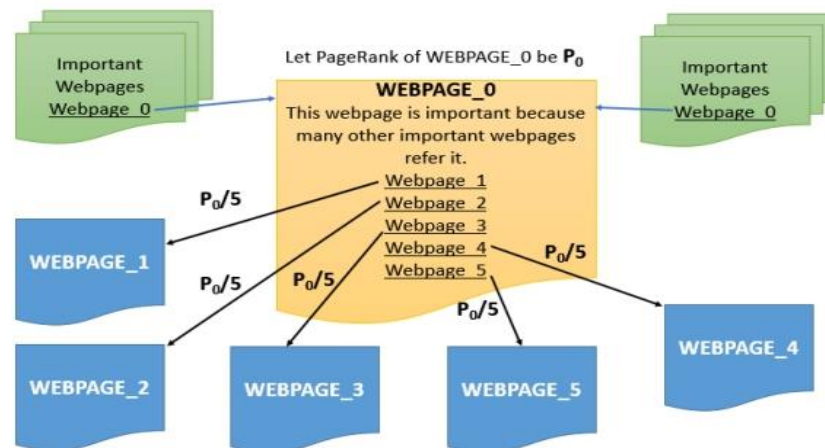
## II. Page Rank Algorithm Illustration

Traditional web search was based on simple keyword-based matching algorithms. It used to follow Boolean model for matching i.e. if exact keyword is contained in a web document then that document should be included in the result set of query. While, the other documents having no exact match were used to be discarded. As the size of web kept on increasing, searching by keywords contained in webpages introduced higher time and space complexities. To overcome this issue, concept of inverted indexes was used in which inverted index contained a list of keywords and corresponding to each keyword a list of pages containing them. So, instead of tracking keywords per webpage, focus shifted towards tracking webpages corresponding to keywords. This section illustrates the Page Rank algorithm.

Most popular search engine, Google uses over a hundred metrics for ranking and returning relevant web results, of which two significant ones (in original implementation) were content score and popularity score [10]. Content score metric was assigned based on matching of keywords present in search query with terms in title or body of web pages. For faster query response times, inverted file structure was used where keywords were indexed along with the web pages that contain these. It improves the processing times over indexing of web pages along with terms they contain.

The second metric was popularity index which was assumed to be the crux of Page Rank method. It was based on importance of web pages. [14] claimed that a web page is important if a large number of incoming links from important pages are referring it. In other words, many important web documents are pointing to this specific web page making it popular. Also if an important page is referring to many other pages, its Page Rank should be distributed proportionately to all those referred web pages. For example, NASA is an important website and suppose it is having out links to 10 other web pages then any of those pages will be distributed one tenth (1/10) of the page rank assigned to NASA website. Figure 1 illustrates this assignment as shown by five outward links from "Webpage 0".

Figure 1: Illustrating PageRank Algorithm [14]



### III. Incorporation of Semantic Technologies into Web Search

During initial advances towards semantic search, applications were prototyped to include semantic elements but without disturbing the prevalent keyword-based search. Semantic Web Search Engine (SWSE) architecture was proposed and prototyped using Jena framework to provide more relevant results [5]. The input given was plain text search queries similar to what is given to traditional search engines. Then utilizing the semantics present in RDF documents, search query keywords were matched against `rdfs:label` and `rdfs:comment` of all available RDF documents resulting into a set of URIs (Uniform Resource Identifier) containing properties and resources (weighted based on frequency of appearance). Then, RDF queries were formed by combining URIs of resources and properties identified. New RDF statements were formed using the new resources detected in this process. Finally, all the RDF statements collected and utilized as semantic graph were matched against input query to yield semantic results based on graph traversal.

#### A. Augmentation of Semantically relevant Information

Web search is being improved by incorporation of semantic web technologies both by augmenting relevant information as well as by increasing the precision of resulting ranked web documents. The original Semantic Search application was prototyped to augment semantically relevant data along with web documents retrieved [7]. Authors explicitly limited the applicability of their application to research search, i.e. "search by concept" or the web information retrieval approach in which search terms convey some concept about which user wish to research. For researching, meaning of the terms should be known along with the interrelationships these have with other concepts. The other approach, which is widely used in traditional information retrieval is navigational search or "search by keyword" in which user just wishes the web documents containing those keywords to be retrieved.

Semantic Search application's main objective is to generate more relevant/accurate results. Accuracy is generally measured using IR metrics of precision, recall and f measure. To reach this goal, machines should be made intelligent enough to understand user's context and intent. For some search queries, we may need additional parameters to disambiguate context and returning highly relevant results.

#### B. Approaches with various technologies

##### 1) Using Semantic Web Technologies

Semantic Web technologies such as Ontology, RDF etc. are being exploited heavily as knowledge bases for deriving relevancy of terms and retrieving precise results from the web. Although, due to lack of such

knowledge repositories encompassing the whole web, large scale IR still require other technologies such as NLP techniques to work in conjunction with semantic web technologies for optimal results. But, for smaller subsets of web, i.e. in restricted domains, approaches using these technologies suffice to retrieve concerned knowledge. Ontology contains machine readable description of a domain without any natural language description. I.e. it contains full, explicit vocabulary to express semantics of a domain for common understanding. Knowledge contained in ontology is meant to be shared and reused [13]. Ontologies may be further categorized as [17]:

- Domain Ontology: Ontology pertaining to a specific domain, containing vocabulary specific and valid to that domain only.
- Generic Ontology: Generic ontology can be spanning across domains due to generic vocabulary.
- Application Ontology: These contain knowledge for modelling a domain. It contains domain knowledge (pertaining to domain to be modelled) and method ontologies.
- Representational Ontology: These just provide representational entities without explicitly mentioning what should be represented. Generally, these does not contain domain knowledge.

Some famous ontology used extensively in literature are:

- WordNet Ontology: WordNet ontology pertains to lexical domain, where English words such as nouns, verbs are organized into synonym set denoting some lexical concept.
- DBpedia Ontology: DBpedia ontology contains domain knowledge across real world concepts available in Wikipedia, the free encyclopaedia.

Some approaches that use ontology for increasing efficiency in information retrieval are:

To overcome the limitation of lower precision as a result of mismatches with exact keywords [16] proposed a novel approach using fuzzy matching instead of exact matching of keywords in books domain. Fuzzy search is used to find matches to keywords approximately. Hence, there are improved chances to match documents that have approximately the query keywords. E.g. for query keywords “International Journal”, documents containing “Int. Journal” may be considered while using fuzzy search, which will otherwise be completely discarded using exact keyword matching search. For implementation, they parsed a collection of 2000 documents about books from Wikipedia and stored relevant fields such as author, publisher, and date etc. in relational database. They built a Book Ontology and populated it with the data stored in relational database. Finally, SPARQL queries were used to retrieve result documents and were compared with the results of traditional (exact) keyword search. Based on the results derived out of 10 queries, their implementation claims an increase of 78% in precision and 13% in recall values over traditional search.

[1] proposed “YaSemIR”: Yet another Semantic Information Retrieval System. Their system used Lucene<sup>2</sup> based standard indexing and a Knowledge Battery containing one or more OWL (Web Ontology Language) ontology and SKOS (Simple Knowledge Organization System) terminology file for denotation of concepts and later annotating these terms when found in text. System implementation annotates ontology concepts found in text and based on these annotations a semantic index is created and maintained separately by expanding the annotations using ontology. For ranking of results they combine metrics for keyword similarity (using standard index) and semantic similarity (using semantic index) into a final score.

There are other approaches concerned with development of specialized ontologies to facilitate efficient retrieval of specific information in specialized domains over web. Ghanem et. al. proposed an ontology for academic research containing vocabulary pertaining to academic publishing [4]. Their approach suggests to populate this ontology using publication information available at websites of international journal publishers such as Springer. Core objective of designing such an ontology is to help academicians find already published work with similar content, which will help them facilitate research and plagiarism detection. Users may query resulting knowledge base to find already published work in the concerned field.

## 2) Using NLP Techniques

NLP techniques can be used for deriving context out of Natural Language (NL) queries posed by users. These can further be used for having an understanding of the web page contents (for ranking relevant web pages). Stanford NLP package is an open source NLP library developed at Stanford University [12]. Wide variety of NLP operations provided through APIs makes it applicable in diverse areas where understanding of sentences is required. Some annotator capabilities available in this package, which are used extensively in literature for performing specific tasks involving NL queries, are tabulated in Table 2.

**Table 2: Annotator capabilities available in Stanford CoreNLP Package**

| Annotator | Description                                                                                                                                                                  |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| TOKENIZE  | Splits a text body into sentence of tokens.                                                                                                                                  |
| POS       | Part of Speech (POS) tagging for tokens. A POS indicator is appended at the end of each token. It considers sentence body to tag appropriate POS in case of ambiguous words. |

|           |                                                                                                                          |
|-----------|--------------------------------------------------------------------------------------------------------------------------|
| LEMMA     | Provides the stem words for tokens.                                                                                      |
| NER       | Named Entity Recognizer (NER) recognizes Person, Date, Money, Location etc. in given sentence.                           |
| SENTIMENT | Sentiment analysis by assigning sentiment scores. Overall sentence could have a positive, negative or neutral sentiment. |
| PARSE     | Full syntactic analysis of sentences.                                                                                    |

[15] made a comparison among various NLP toolkits available. They compared performance of standard toolkits on formal texts as well as social media texts. Performance tests performed were for Tokenization, POS Tagging, Chunking and Named Entity Recognition on standard datasets. Standard toolkits compared were NLTK, Apache OpenNLP, Stanford CoreNLP and Pattern Toolkit. Performance measured using standard metrics of Precision, Recall and F-measure concluded that Apache OpenNLP Library is yielding highest performance on formal texts. Though some other toolkits outperformed others in specific tests.

### 3) Semantic Search Engines

Dong et. al. categorized semantic search engines into six classes. These are tabulated in Table 3 along with their characteristics.

**Table 3: Classification of Semantic Search Engines [2]**

| Classified as:                     | Characteristics                                                                                                                                                                                                                                                                                                                                     |
|------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Semantic Search Engines            | <ul style="list-style-type: none"> <li>Purely based on semantic web technologies</li> <li>Efficient for restricted domains,</li> <li>May be inefficient for large scale search environments (such as the Web)</li> </ul>                                                                                                                            |
| Semantic Search Methods            | <ul style="list-style-type: none"> <li>Usage of various semantic technologies such as Description Logics, DAML+OIL, RDF and semantic annotation for improving search</li> </ul>                                                                                                                                                                     |
| Hybrid Semantic Search Systems     | <ul style="list-style-type: none"> <li>Using semantic web technologies with traditional keyword based text search engines for improving IR metrics</li> </ul>                                                                                                                                                                                       |
| XML Search Engines                 | <ul style="list-style-type: none"> <li>Search in structured documents (XML) for concepts</li> </ul>                                                                                                                                                                                                                                                 |
| Ontology Search Engines            | <ul style="list-style-type: none"> <li>Search appropriate domain ontologies by means of keyword based search where keywords are concepts</li> <li>Filtering options for restricting domain of ontologies to find best fit</li> </ul>                                                                                                                |
| Semantic Multimedia Search Engines | <ul style="list-style-type: none"> <li>Semantically search and retrieve concerned multimedia files (such as audio and images)</li> <li>Use advanced technologies such as Speech-to-Text conversions, Computer Vision etc. to capture textual metadata</li> <li>Semantic search over textual metadata to return relevant multimedia files</li> </ul> |

Due to various limitations on large scale web search (as listed in Table 1), universal semantic search engines are yet to be made efficient in terms of relevancy and precision of results. Although, in restricted domains where there is little ambiguity in vocabulary and concepts are well defined, a number of semantic search engines are proposed and implemented. Currently, these domains are restricted to life sciences and biomedical research.

[8] proposed BioSearch, a Biomedical domain semantic search engine that utilize ontologies to assist query formulation and result organization. Built using Java and some browser side JavaScript, its user interface assist users to filter query terms according to datasets and concepts. For user evaluation, queries were formulated and fired on eight selected Bio2RDF linked datasets, resulting in better efficiency and usability than its comparison systems.

[3] implemented SEMEDICO, a semantic search engine in life sciences domain for querying literature in that domain. It is available as an Apache Tapestry Web application. Some of its unique features include assistive query formulation and interactive concept disambiguation. It heavily utilizes NLP techniques for named entities recognition and species tagging.

## IV. Semantic Knowledge Presentation Schemes

At present, incorporation of semantic elements in information retrieval mechanisms over the web can be best seen in the way results are presented by most of the popular search engines such as Google and Bing. Apart from displaying list of matching web documents corresponding to search query, other schemes such as Related Searches, Recommendations, Semantic Auto-completes, and Relevant Targeted Advertisements are also often presented. These suggestions/recommendations are purely based on the semantic relevance to the search query.

### A. Recommendation Sidebars

Recommendations are the most visible application of semantics based searching, which can be seen along with result set of search query. Figure 2 illustrates the recommended searches provided by Google search engine for a search query "Parachute". The key point here is that although, "Glider", "Kite", "Airship" etc. have no syntactic relationship with "Parachute" but semantically, they denote the instances of a class denoting "Flying Objects". These recommendations are often provided based on:

- Current User's search history
- Cache of similar searches by other users (Trending Searches)

**Figure 2: Recommendations presented on searching for “parachute” by Google<sup>3</sup>**



### **B. “Did you mean” feature**

Auto complete feature was introduced way back in 2004 to provide suggestions to keyword being typed. It was based on the frequency of keywords being queried earlier containing prefix as the partial keyword being typed. Often Ternary Search Tree (TST) data structure is used for implementing spell corrections. Also, search engines often use adaptive mechanisms to correct spellings with user feedback over time (E.g. user doesn’t prefer to visit links which resulted from querying misspelt words) [6].

### **C. Wikipedia Infobox**

Wikipedia, the free encyclopedia is the fifth most visited website on the web<sup>4</sup>. Its popularity is due to availability of massive trusted information it contains in the form of millions of articles about people, things, locations etc. Being a host of such massive knowledge repository which is freely available, some projects were initialized to utilize and extract the structured knowledge out of its huge text collections. One such project is Wikipedia Infobox<sup>5</sup>, which aimed to present relevant structured information as a concise summary of key facts about a subject. Knowledge is represented in a template form very similar to that of triples. There are certain benefits of knowledge represented in infoboxes<sup>5</sup>:

- It contains Machine readable metadata (structured content) which may be made available for other tools like DBpedia and Freebase later.
- It provides a concise and small summary of key facts about a subject
- Article editors do not need to know anything about metadata standards or markup (a way to structure unstructured data without changing the way addition of unstructured data happens).

Also, to provide uniformity across infoboxes, some key directions are given for standardization of content contained in them:

- Designers should create and use standard set of parameters to be used in related infoboxes
- It is recommended to use same names for same parameters across infoboxes (website vs. homepage vs. URL are all same). This is also illustrated in Figure 3 where Java and C++ being two popular programming languages, share almost the same template for triples such as paradigm, designed by, type discipline, license etc.
- It is highly desirable to merge redundant/duplicate infoboxes.

**Figure 3: Wikipedia Infoboxes for Java and C++ showing similar templates (highlighted)** <sup>6,7</sup>

|                                                                                   |                                                                                                        |                                                                                    |                                                                                 |
|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
|  |                                                                                                        |  |                                                                                 |
| <b>Paradigm</b>                                                                   | Multi-paradigm: Object-oriented (class-based), structured, imperative, generic, reflective, concurrent | <b>Paradigm</b>                                                                    | Multi-paradigm: procedural, functional, object-oriented, generic <sup>[1]</sup> |
| <b>Designed by</b>                                                                | James Gosling                                                                                          | <b>Designed by</b>                                                                 | Bjarne Stroustrup                                                               |
| <b>Developer</b>                                                                  | Sun Microsystems (now owned by Oracle Corporation)                                                     | <b>First appeared</b>                                                              | 1985; 33 years ago                                                              |
| <b>First appeared</b>                                                             | May 23, 1995; 22 years ago <sup>[1]</sup>                                                              | <b>Stable release</b>                                                              | ISO/IEC 14882:2017 / 1 December 2017; 3 months ago                              |
| <b>Typing discipline</b>                                                          | Static, strong, safe, nominative, manifest                                                             | <b>Typing discipline</b>                                                           | Static, nominative, partially inferred                                          |
| <b>License</b>                                                                    | GNU General Public License, Java Community Process                                                     | <b>Implementation language</b>                                                     | C++ or C                                                                        |
| <b>Filename extensions</b>                                                        | .java, .class, .jar                                                                                    | <b>Filename extensions</b>                                                         | .C, .cc, .cpp, .cxx, .c++, .h, .hh, .hpp, .hxx, .h++                            |
| <b>Website</b>                                                                    | oracle.com/java/                                                                                       | <b>Website</b>                                                                     | isocpp.org                                                                      |

## V. Discussion

Due to exponentially increasing size of web, keyword-based IR techniques often fail to retrieve precise results. Thus, we are in a phase to adopt semantics-based IR techniques using key semantic web technologies combined with NLP techniques. Though, these semantics-based IR techniques are being successful in small domains with restricted domain vocabulary, but their impact in full scale web search is still limited due to multiple factors. These factors include collisions among domain vocabularies (i.e. same term being used in different domains with often very different meanings), unavailability of massive knowledge bases (though, works like DBPedia and Google Knowledge Base are still progressing) and efficient mechanisms for measuring user satisfaction on result presentation.

But at present, utilization of semantics along with traditional IR is certainly improving quality of retrieved results as is best seen with highly relevant recommendations and suggestions along with search results. For facilitating this, concepts from semantic knowledge repositories are augmented with query keywords to expand the search query and get relevant results. Query composed of conjunction operator (AND) becomes more specific, whereas composed of disjunction operator (OR) becomes more general. Long conjunctive queries often yield high precision as we are filtering out results and making space only for precise results whereas, long disjunctive queries often yield high recall as we are expecting result every combination of query terms to be present in our result set [11].

Finally, to see the impact of semantics-based technologies in search, let us consider three categories of search engines as highlighted by [9]:

- General search engines: These are used to query whole of the web.
- Specific Domain search engines: These contain (index) documents pertaining to specific domains.
- User profiling based search engines: These contain past user search history and search behaviours with provision for user to reject some irrelevant personalized results.

Semantic search for General search engines is still in evolving phase and efforts are needed to make IR much efficient. Also, schemes like recommendation, semantic suggestions etc. are showing steady progress in this direction. Restricted domain systems have best IR metrics of precision and recall due to availability of precise concepts in its vocabulary as well as comparably smaller footprint of documents to search. These metrics are often benefited by usage of domain specific ontologies for semantic search within that domain. User profiling based search systems are the best sources of recommendations, as they have access to user's search history as well as their preferences. Further, with incorporation of a feedback path of user acceptance or rejection of results, personalized results may become highly relevant over time using adaptive techniques like machine learning.

## VI. Conclusion and Future Work

IR over web approaches are steadily moving towards the usage of semantics in contrast to usage of keyword-based similarity as the only criteria. Due to enormous size of web, this transition has to be slow without disturbing the existing keyword-based search. Hence, the semantics-based approaches are being augmented to

traditional IR instead of replacing it. Some traditional approaches such as Page Rank has been illustrated. Then, various advances in search over the web are explored from a semantic IR perspective. Approaches explored utilizes semantic web technologies in conjunction with NLP techniques for improving efficiency of result set over web. Applications of these approaches can be best seen in result presentation schemes provided by most of the popular search engines today. Recommendation sidebars and Wikipedia Infoboxes are among such schemes. As part of the future scope, this work can be further expanded by exploring and adding a timeline of semantic search advancements since the vision of semantic web and its technologies. Also, overlapping of related disciplines such as Artificial Intelligence (AI), Machine Learning etc. with Semantic IR can be studied, considering real world applications of research in this direction such as Google TalkToBooks<sup>8</sup>, a project exploiting AI, IR and NLP techniques, letting users explore relevant book passages just by issuing NL queries.

## VII. References

- [1] D. Buscaldi and H. Zargayouna, "YaSemIR: Yet another Semantic Information Retrieval System," ESAIR 2013, Oct 2013, San Francisco, United States. pp.13-16, 2013.
- [2] H. Dong, F. K. Hussain, & E. Chang, "A survey in semantic search technologies," In *Digital Ecosystems and Technologies, Feb 2008. DEST 2008. 2nd IEEE International Conference on* (pp. 403-408). IEEE.
- [3] E. Faessler and U. Hahn, "SEMEDICO: A Comprehensive Semantic Search Engine for the Life Sciences," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 91–96, Vancouver, Canada, July 30 - August 4, 2017.
- [4] M. Ghanem, A. Mouloudi, and M. Mourchid, "Towards a scientific research based on semantic web," The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015), Procedia Computer Science 73 (2015) 328 – 335.
- [5] D. E. Goldschmidt and M. Krishnamoorthy, "Architecting a search engine for the semantic web," In: Proc. of the AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications, 2005.
- [6] [http://www.googleguide.com/spelling\\_corrections.html](http://www.googleguide.com/spelling_corrections.html)
- [7] R. Guha, R. McCool and E. Miller, "Semantic Search,," WWW2003, May 20-24, 2003, Budapest, Hungary. ACM 1-58113-680-3/03/0005
- [8] W. Hu, H. Qiu., J. Huang et al. "BioSearch: a semantic search engine for Bio2RDF," Database (2017) Vol. 2017: article ID bax059; doi:10.1093/database/bax059
- [9] M. Klopotek, "Intelligent Information Retrieval on the Web," In: Szczepaniak P.S., Segovia J., Kacprzyk J., Zadeh L.A. (eds) Intelligent Exploration of the Web. Studies in Fuzziness and Soft Computing, vol 111. Physica, Heidelberg, 2003.
- [10] A. N. Langville and C. D. Meyer, "Information Retrieval and Web Search," In: L. Hogben, R. Brualdi, A. Greenbaum, and R. Mathias, eds., Handbook of Linear Algebra, chapter 63, 63–1, CRC Press, 2006.
- [11] C. Mangold, 'A survey and classification of semantic search approaches', Int. J. Metadata, Semantics and Ontology, Vol. 2, No. 1, pp.23–34, 2007.
- [12] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. Mc-Closky, "The stanford corenlp natural language processing toolkit," In Proceedings of 52<sup>nd</sup> Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, 2014.
- [13] B. Omelayenko, "Machine learning for ontology learning," Presentation at the Int. Jyväskylä Summer School, Finland, 2000.
- [14] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Technical report, Stanford Digital Library Technologies project, 1998.
- [15] A. Pinto, H. Gonçalves Oliveira and A. A. Oliveira, "Comparing the performance of different NLP toolkits in formal and social media text," In OASIS-OpenAccess Series in Informatics (Vol. 51). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.
- [16] S. Remi and S. C. Varghse, "Domain ontology driven fuzzy semantic information retrieval," Procedia Computer Science 46, 676–681, 2015.
- [17] R. Studer, R. Benjamins and D. Fensel, "Knowledge Engineering: Principles and methods," Data and Knowledge Engineering, 25: 161-197, 1998.

### Footnotes:

<sup>1</sup> <http://www.searchenginehistory.com/>

<sup>2</sup> <http://lucene.apache.org>

<sup>3</sup> <https://www.google.co.in/search?q=parachute>

<sup>4</sup> <https://www.alexa.com/topsites>

<sup>5</sup> [https://en.wikipedia.org/wiki/Wikipedia:WikiProject\\_Infoboxes](https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Infoboxes)

<sup>6</sup> [https://en.wikipedia.org/wiki/Java\\_\(programming\\_language\)](https://en.wikipedia.org/wiki/Java_(programming_language))

<sup>7</sup> <https://en.wikipedia.org/wiki/C%2B%2B>

<sup>8</sup> <https://books.google.com/talktobooks/>