An Semantic Approach using Query Expansion

Harmeet Singh Lamba[1] Prajakta Subhash Patil[2] Dilip Maheshchandra Nayak[3] Prof. Shweta Barshe[4]

[1,2,3]B.E. Student [4]Assistant Professor

[1,2,3,4]Department of Computer Engineering

[1,2,3,4]Bharati Vidyapeeth College of Engineering, Navi Mumbai-400614

Abstract— The growing amount of information increases the size of the databases. To search the required information from these big data repositories we need some specially designed search techniques for a specific website. There are many searching techniques, but retrieving relevant information is difficult. To overcome these difficulties, semantic web technologies are playing a vital role. However, expansion terms are usually determined on "term co-occurrences" within records. This paper describes a method of implementing a semantic search using query expansion as well as retrieving relevant information by content ranking. These records are retrieved using query expansion and then they were ranked to provide users with the most relevant information. This proposed method helps the information retrieved maintain its relevance.

Key words: Semantic Search, Various Semantic Search approaches, Query Expansion, Approaches of Query Expansion

I. INTRODUCTION

With the growth of internet, there are a growth of many easy ways of accessing information and services. Web provides vast amount of information, with the increasing information we are facing new problem for locating relevant information. Efficient searching is required to acquire high quality relevant result. When the user uses search fields to search for specific information, the quality of search result will be improved significantly if they make use of advanced techniques. Most of the traditional search approaches get the results syntactically correct but large in amount.

The semantic allows the information to be precisely described in terms of well-defined vocabulary. Semantic Web is gaining momentum. A semantic search gives selected results which the user is searching for. The main objective of Semantic Web is to make Web content understandable not only by humans, but also machine understandable. We need to ensure that semantics are not lost during the whole life cycle of information retrieval. Which can be discussed in detailed in later sections.

```
SELECT * FROM Event
WHERE Tense='PRESENT' OR Tense='UPCOMING'
ORDER BY Count, Event_Name;
```

For example in a system, to provide a motivated value of the "Count" - attribute for each tuple. The rest of the paper is organized as follows. First, we explain the limitations of traditional search and then how semantic search overcomes those drawbacks, then we also portrait the different approaches by which we can achieve semantics in search. Then we explain the query expansion technique which we have used to understand user's intent, and finally by content ranking we try to achieve relevant results that user prefer more. This relevance is judged on the basis of previous visits to the records. The visits are tracked by us and then displayed by highest visit first order.

II. TRADITIONAL SEARCH

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. Current traditional search technologies have reached a plateau. For our traditional search engines, the "amount of Web content outpaces 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. 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 are:

A. Limitations of Traditional Search:

1) Problem due to Polysemy words (one word having several meaning).e.g. word "Bank" it can be financial institution or river shore.
2) 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 Baby" is a song title, and the meaning of "baby" in this 131 entity is different than the usual meaning of "infant".
3) Traditional Information Retrieval (IR) technology is based almost purely on the occurrence of words in records. 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.
4) Low Precision and Low Recall Problem. Precision is the fraction of the records retrieved that are relevant to the user's information need. While recall is the fraction of the records that are relevant to the query that are successfully retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

These limitations are overcome by Semantic Search Techniques that brought a revolutionary change in the search Approaches.

III. SEMANTIC SEARCH

The Semantic Search is an extension of the traditional search where information is represented in a machine process able way. While the information on the Web is mostly represented as HTML documents. The semantic search highly improves search accuracy of the query related data and the search algorithm delivers the exact content, the user intent to know. Semantic search seeks to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data space, whether on the Web or within a closed system, to generate more relevant results. There are other ways to search the web, using semantic search engines. Semantic search systems consider various points including context of search, location, intent, variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries to provide relevant search results.

A semantic search 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 Semantic Web and search tool to improve the search results gained by current search tools and evolves to next generation of search engines built on Semantic Web.

By using semantic search we will ensure that it results in more relevant and smart as required. The search engines are able to compare or extract the data and gives very relevant results for the queries. Major web search engines like Google and Bing incorporate some elements of semantic search. The various approaches or methods that could be used for semantic search engines are explained in the following section.

IV. VARIOUS SEMANTIC SEARCH APPROACHES

Semantic search Includes Four main Approaches. Different semantic search engines may use one or more of these approaches. The point of semantic search is to use searched keyword meaning to improve the user's search experience. For example,

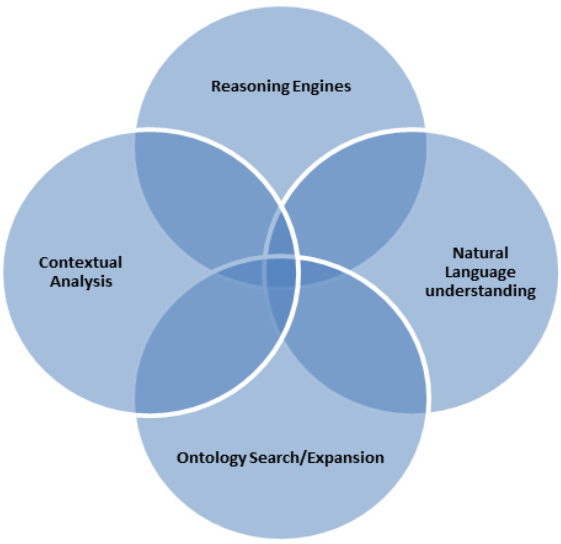

Fig. 1: Approaches to Semantic Search

First approach uses contextual analysis to help to disambiguate queries. For example, the word "strike," refers to baseball or labor or something else entirely.

Second approach is focused reasoning. Given set of facts that are represented in the system, additional facts can be inferred from them. If the system knows who is Bach's children were, and it knows who each of their children were, then a reasoning system can infer who is Bach's grand children were.

Third approach Point out on natural language understanding. These engines process the content they index and the queries people submit try to identify the intent of the information. In this approach they use the syntax of the sentence and rules to identify people, places, organizations, and so forth. Power set makes voluminous use of natural language understanding.

The fourth approach uses ontology to perform knowledge about a domain and expanded queries. In this approach, when a user enters a query for a word like "Car," the system adds terms from its ontology (e.g., "vehicle" because a Car is a kind of vehicle) to make the search more focused and accurate as well as more broad This approach is used by a large number of semantic search systems.

V. QUERY EXPANSION

Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. In the context of web search engines, query expansion involves evaluating a user's input (what words were typed into the search query area and sometimes other types of data) and expanding the search query to match additional records.

The problem for concept matching is that many records semantically related to the query do not necessarily contain the query term. So the retrieval model is not able to find semantically related records, if they do not contain the query term. Query expansion by adding related terms to the query is a common technique to handle this problem. The expanded queries enable the retrieval of additional records that don't contain the query term but are semantically related to the query.

Query expansion involves techniques such as:

- Finding synonyms of words, and searching for the synonyms as well
- Finding all the various morphological forms of words by stemming each word in the search query
- Fixing spelling errors and automatically searching for the corrected form or suggesting it in the results
- Re-weighting the terms in the original query

Query expansion is a methodology studied in the field of computer science, particularly within the area of natural language processing and information retrieval.

Query expansion techniques have been studied for nearly three decades. The various methods proposed in the literature can be classified into the following three groups: query specific, corpus specific, and language specific.

(a) Query-specific terms can be identified by locating new terms in a subset of the result set retrieved by a specific query. This is the approach taken by relevance feedback systems, where related terms come from the contents of user-identified relevant information. This has been shown to be quite effective, but it requires that users indicate which records are relevant. More recently, search improvements are being achieved without the user's relevance judgments.

(b) Corpus-specific terms are found by analyzing the contents of a particular full-text database to identify terms used in similar ways. It may be hand-built, a time-consuming and ad hoc process, or created automatically. Traditional automatic thesaurus construction techniques group words together based on their occurrence patterns at a document level, that is, words which often occur together in text are assumed to be similar. These thesauri can then be used for automatic or manual query expansion.

(c) Language-specific terms may be found from generally available online thesauri that are not tailored for any particular text collection.

This adopts an automatic query-specific terms approach for locating related terms. We are particularly interested in these techniques because they are commonly used to add useful words to a query. Unfortunately, casual users seldom provide a system with the relevance judgments needed in relevance feedback. In such situations, ad hoc or blind feedback is commonly used to expand the user query. This method takes the form of pseudo-relevance feedback, where the actual input from the user is not required. In this method, a small set of records is retrieved using the original user query; these records are all assumed to be relevant without any intervention by the user. The content of the assessed records is used to adjust the weights of terms in the original query and/or to add keywords to the query. The new query is reformulated towards relevant records and away from the non-relevant ones.

VI. APPROACHES TO QUERY EXPANSION

Some approaches to query expansion are stated in this section. The existing state-of-the-art query expansion approaches can be classified mainly into two categories — techniques based on global analysis, which obtains expansion terms on the statistics of terms in the whole corpus, and local analysis, which extracts expansion terms from a subset of the search results.

Now typically a query analyzer will receive query fired from the browser. Query analyzer will split the query into multiple sub queries as follows:

So if the query is like "Google Mountain view jobs" Subqueries:

1) Complete query: "Google Mountain view jobs"
2) Boolean query: "Google AND Mountain AND view AND jobs"
3) Distance / Proximity query: (Google Mountain) AND (Mountain View) AND (View jobs) AND (Google Mountain View) AND (Mountain View Jobs)
4) Single word queries: Google, Mountain, view, Jobs
5) Single word fuzzy queries: Google~, Mountain~, view~, jobs~
6) Range Query: if the search term includes a date range, regular expressions will pick up the date range and create a date range query.

A. Global Analysis:

In this section, we only review the approaches that exploit term co-occurrences in records. We do not analyze the approaches that use a manual thesaurus (e.g., WordNet) One can refer to for some examples of utilization of such a resource for query expansion. Global analysis is one of the first techniques to produce consistent and effective improvements through query expansion. The basic idea of global analysis is to use the context of a term to determine its similarity with other terms. Global analysis selects expansion terms on the basis of the information on the whole record set. It builds a set of statistical term relationships which are then used to expand queries. One of the earliest global analysis techniques is term clustering. Queries are simply expanded by adding similar terms that are grouped into the same cluster according to term co-occurrences in records. PhraseFinder is a component of the INQUERY system that creates an association thesaurus. The phrases selected by PhraseFinder are used in query expansion. Latent Semantic Indexing can also be viewed as a kind of query expansion. In its reduced dimensional space, implicit correlations among terms can be discovered and employed in expanding original queries. Generally, global analysis requires corpus-wide statistics, such as statistics of co-occurrences of pairs of terms, resulting in a matrix of similarities between terms or a global association thesaurus. Although the global analysis techniques are relatively robust, the corpus-wide statistical analysis consumes a considerable amount of computing resources. Moreover, since it focuses only on the record side and does not take into account the query side, global analysis only provides a partial solution to the term mismatching problem.

B. Local Analysis:

Different from global analysis, local analysis uses only a subset of records that is returned with the given query. The result is thus more focused on the given query than global analysis. Local analysis techniques are grouped into two categories: approaches based on user feedback information and approaches based on information derived from a subset of the returned records.

VII. INFORMATION RETRIEVAL

IR is an acronym for Information Retrieval. It embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, or machines are employed to carry out the operation. Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs. User queries are matched against the database information. Depending on the application the data objects may be, for example, text documents, images, audio, mind maps or videos. Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

A. Traditional Information Retrieval:

It has two streams of activities. First, one is the systems side with processes performed by the system and other is the user side with processes performed by users. These two sides led to "system orientation" & "user orientation". In system side automatic processing is done; in user side human processing is done. They meet at the matching process where the query is fed into the system and system looks for records that

match the query Also feedback is involved so that things change based on results e.g. query is modified & new matching done

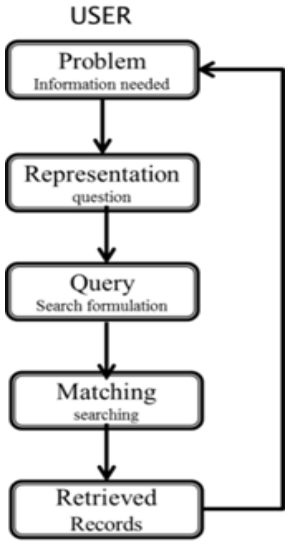

Fig. 2: Traditional Informational Retrieval

The steps shown above describe the flow of information retrieval process. First in the problem step the user specifies the information required. In Representation step it is formulated as question which then is queried as a whole keyword to the database. These keywords are then matched and the matching records are retrieved as a result to the user.

But the information retrieved is in random fashion it must be filtered by some technique that provides the relevancy thus Relevant Information Retrieval could be used

B. Relevant Information Retrieval:

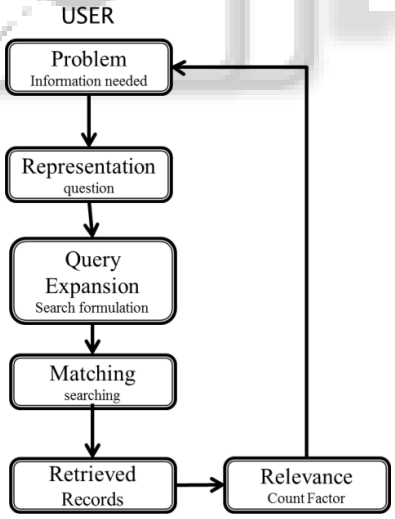

Fig. 3: Relevant Information Retrieval

The steps shown above describe the flow of information retrieval process. First in the problem step the user specifies the information required. In Representation step it is formulated as question. This question is broken down into individual keywords. Then these keywords are used in a query separately. Then each query is fired on database one after the other creating intermediate results. The results are stored in a result set with co-occurrence of all keywords. These keywords are then matched and the matching records are retrieved as a result to the user.

Each time a user visits (clicks) a record a count is incremented within the database. The next time database is queried it orders the result set on the basis of highest visits first order.

VIII. Conclusions

In this paper we have provided a brief overview of some of the best semantic search engines that uses various approaches in different ways to yield unique search experience for users. We have also provided detail explanation about query expansion and its approaches. We used these technologies to provide user with relevant information. It is concluded that searching the internet today is challenge and it is estimated that nearly half of the lex questions go unanswered. Semantic search has the power to enhance the traditional web search. Whether a search approach can meet all these criteria continues to remain a question. Future enhancements include developing an efficient semantic web search technology that should meet the challenges efficiently and compatibility with global standards of web technology.

IX. Acknowledgement

Sincere appreciation and warmest thanks are extended to the many individuals who is in their own ways have inspired us in the completion of this project.

Firstly we are thankful to our principal DR. M. Z. Shaikh for his help. We are extremely grateful for his friendly support and professionalism. We express our heartfelt gratitude to our Head of Department Prof. D. R. Ingle & project coordinator Prof. Rahul Patil of Computer Department for their help and support. This task would have not been possible without the help and guidance of our esteemed project supervisor Prof. Shweta Barshe, without her expert help and guidance, this project would not have reached this stage. We are also conveying special thanks to all staff members of Computer Engineering Department for their support and help. Last but not least, we are very much thankful to our friends who directly or indirectly helped us in completion of the project report.

Reference

[1] G.Sudeepthi, G. Anuradha, M.Surendra Prasad Babu,"A Survey on Semantic Web Search Engine", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.

[2] G. Madhu, Dr.A.Govardhan, Dr.T.V.Rajinikanth, "Intelligent Semantic Web Search Engines: A Brief Survey", International journal of Web & Semantic Technology (IJWesT) Vol.2, No.1, January 2011.

[3] David Parry, "'Tell Me The Important Stuff' - Fuzzy Ontologies And Personal Assessments For Interaction With The Semantic Web", December 1, 2007.

[4] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma, "Query Expansion by Mining User Logs", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, NO. 4, July/August 2003.

[5] Mark Sanderson and W. Bruce Croft, "The History of Information Retrieval Research", Vol. 100, May 13th, 2012.

[6] WANG Peng (汪 鹏), XU Baowen (徐宝文), ZHOU Yuming (周毓明), "Extracting Semantic Subgraphs to Capture the Real Meanings of Ontology Elements", Volume 15, Number 6, December 2010.
[7] Ronald R. Yager, Rachel L. Yager, "Soft Retrieval and Uncertain Databases", 47th Hawaii International Conference on System Science 2014.
[8] T. Finin, J. Mayfield, C. Fink, A. Joshi, and R. S. Cost, "Information retrieval and the semantic web," in Proceedings of the 38th International Conference on System Sciences, Hawaii, United States of America, 2005.
[9] R. Guha, R. McCool, and E. Miller, "Semantic search," in Proc. of the 12th international conference on World Wide Web, New Orleans, 2003, pp. 700–709.
[10] Arthur H. van Bunningen, Maarten M. Fokkinga, Peter M.G. Apers, "Ranking Query Results using Context-Aware Preferences", Data Engineering Workshop, 2007 IEEE 23rd International Conference on, 17-20 April 2007, 269 – 276.
[11] Mouna Kacimi, Fabian M. Suchanek, Aparna Varde, "Databases, Information Retrieval and Knowledge Management: Exploring Paths and Crossing Bridges", Vol. 42, No. 3, SIGMOD Record, September 2013.

