Design and Implementation of Semantic based Search Engine for Punjabi

Tanjyot Aurora
M.Tech Scholar Department of Computer Engineering, Punjabi University, Patiala.

Brahmaleen Kaur
Assistant Professor, Department of Computer Engineering,
Punjabi University, Patiala.

ABSTRACT
The World Wide Web (WWW), helps to share information globally, the amount information has outgrown billions of databases. Its size, heterogeneity and human –oriented semantics pose a serious obstacle in search for desired information. Traditional search engines like Google help naïve users to access web resources and perform only word based indexing and searching may provide with irrelevant documents that are presented to user. Addition of semantics to search engine solves the two problems associated with search engines low precision and low recall. Work has been done in English and Hindi dialect but not much work has been done for Punjabi dialect. This paper describes the design and working of online semantic based search engine for Punjabi 'UDDAN' and also proposed an algorithm so as the results obtained have high precision.

General Terms
Information Retrieval, Search engine, Semantic based Search engine, World Wide Web (WWW)

Keywords
Punjabi based Semantic Search engine, Query Expansion, Semantic search

1. INTRODUCTION
Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources [1]. Web search engines are the most evident IR applications. The moment user enters query into the system the process of information retrieval begins. Queries are formal proclamations of information needs, for example search inputs in web search engines. In information retrieval a query does not distinguish a single object in the collection. Instead, several objects may match the query, may be with different degrees of pertinence.

An entity is a real world object that is represented by information in a database. User queries are assessed against the database information. Contingent upon the use the data objects may be, for instance, content archives, pictures, sound and features. Frequently the archives themselves are not put away straightforwardly in the IR framework, however are rather stored in framework in metadata form. Most IR systems calculate 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 methodology might then be iterated if the user wishes to refine the query. When the human computer interactive interface was implemented, the query expansion technology started to be used in information retrieval system. It has relation with command language, menu selection; diagram operation, natural language communication and all other information retrieval methods [2] .Queries by users are constricted to operations and knowledge of users.AQuery can be complicated or simple which is dependent on user.

2. SEMANTIC SEARCH
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 [3]. Semantic search systems consider various points including context of search, location, intent and variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries to provide relevant search results. Rather than using ranking algorithms such as Google's Page Rank to predict relevancy, semantic search uses semantics or the science of meaning in language, to produce highly relevant search results. In most cases, the goal is to deliver the information queried by a user rather than have a user sort through a list of loosely related keyword results. The Web is growing as the quickest communication medium. This innovation in mix with most recent electronic stockpiling gadgets, empower us to stay informed concerning tremendous measure of information accessible to the general public. Information present in web is available in diverse dialect like English, Arabic, Bengali, Hindi a lot of people more. The Web is brimming with unstructured data and customary web engines hone a keyword -oriented search which prompts the issue of retrieving various unessential data. Such search technique with a particular keyword may eventuate to unsuitable but considering its synonym to relevant results. A hefty portion of the results recovered for general questions are immaterial to the subject of investigation and different archives are missing in light of the fact that the query does exclude the careful key phrases. users are not certain about the dialects used to plan their inquiries, refined questions with restrictive Boolean operators may bring about a couple of or even no records .This inspires the utilization of regular dialect interfaces as an adequate method for communicating with web engines. There are two major forms of search: navigational and research [4]. In navigational search, the client is utilizing the web engine as a navigation tool to navigate to a particular intended document. Semantic search is not applicable to navigational searches. In research search, the user feeds the search engine with an expression which is planned to mean an item about which the user is attempting to assemble/research data. There is no specific document which the client thinks about and is attempting to get to. Rather, the client is attempting to place various archives which together will give the relevant data.

3. COMPARISON WITH TRADITIONAL SEARCH ENGINES

The traditional search engines they search the web using keyword oriented schemes. It doesn't consider the meanings but just matches the keywords and page is retrieved according to page rank algorithm of the particular search engine. Thus, the results produced are irrelevant and have low accuracy. Search process is affected by semantic webs' features.

The Distinguishing features are listed below [5]:

All objects of real world are involved in the search process.

Knowledge is understandable for machine as well as human.

Semantic web languages are well structured than HTML.

Asingle concept can be represented by distributed knowledge.

These features cause fundamental differences to traditional Search engines. They are listed below [5]:

• Intelligent retrieval is provided by using a logical framework.

• In documents metadata maintenance, update and more complex ranking is resulted by complex relations.

• Visualization techniques are required for visualization of search results by specifying relationships among objects.

4. RELATED WORK

According to research accomplished by Senthil J et.al. [6] Semantics is the study of meaning. It is centralized on relationships like words, phrases, symbols etc. and proposed a methodology for semantic search engine which included handling polysemy, decision inputs ,k-nearest paradigm, decision boundaries, decision trees, Nodes as vectors of tags, and Semantic similarity. When a user inputs a search query, it may have many possible meanings which were represented in form of assumptions. The search results were made more relevant and accurate as compared to usual search engines by further involving key factors such as query generalization and specialization and concept matching and result was a artificially trained search engine which was able to make intelligent inferences. The search engine developed was fact and knowledge oriented. Rather than displaying the web links as results, it presented only the information intended by user. Junaid Mohamed Kassim et.al.[7] stated that a typical web engine consists of three parts which includes:- (1) A database of web documents(2)A search engine operating on that database(3) A series of programs that determine how search results are displayed. Different from traditional search engines, a semantic search stores semantic information about web sources .It combines the technologies of semantic web and search engine to improve the search results gained by traditional search engines. The process consists of following steps:-

• The user query is interpreted, extracting the relevant concepts from sentences

• That set of concepts is used to build a query that is launched against ontology

• Results are presented to user.

Qazi Mudassar IIyas et.al. [8] Emphasized the utility of semantic web which makes it possible to successfully execute the query as it allows for associating formal meaning with content .The main focus of semantic web is to make the web machine understandable where automated agents will be able to understand the content on Web. A layout was proposed for semantic search engine which encompassed the creation of ontologies in plain text format and its translation into database translator. The Crawler would find new Ontologies, then after annotation queries sent to inference engine which then reasons on queries by using database and finally sent to user to be viewed on web. For semantic search engine to become a reality all content on web must be annotated with data which is a major issue yet to be resolved. D.Schneider et.al. [9] addressed the issue to develop a semantics enabled multimedia search engine by representing multimedia content as textual results .It considered the fact that multimedia documents like PowerPoint Presentations or flash documents are widely used in internet and there is no way to explore and search for content .The system was named "Fulgeo". FLAME (Flash Access and Management environment) was the only existing search engine. The new system developed would consider the initial keyword based query and match the text extracted from flash files as well as semantic concepts related ,a thumbnail of the result was shown, such that results with larger thumbnail seem to be more important to user. Nandkishor Vasnik et.al. [11] adopted a method to improve the relevancy of retrieved results by expanding the user query with more relevant words. Sometimes, the domain of query is unpredictable so enhancing keywords is really difficult. The output of search engine is dependent upon the database present and how structured is the database. The proposed query expansion model included three methods for query enhancement: The first method utilized Lexical resource, second method utilized user context and the last method devised was a combination of both. An experiment was conducted with 30 queries in four modes of search which consist of (1)Simple google search,(2)Method-I (Query enhance with HWN)(3)Method-II(Query enhance with user context)(4)Method-III(Expansion with HWN and user context) .Precision values were calculated Considering the uppermost retrieved documents by experiment which showed that a combination of Method-I and Method-II could enhance the information retrieval. Maleerat Sodani et.al. [12] Conducted an experiment for subjective query expansion which considered that the quality of searching is most important. There are numerous strategies to enhance the quality search, the query expansion (QE) is considered to enhance the query terms with regard to fulfil user query requirements .The expansion method utilized the keyword based query. These semantic terms are added to the first query so as to reformulate query before sending it to the searching module. The tests were tried on twitter information accumulated. The results demonstrated that the retrieval adequacy is impressively higher than utilizing just original query. Contrasted with the baseline system, this technique gives higher execution regarding review and accuracy recall and user satisfaction. Aurora et.al. [13] has defined for implementing semantic based search engine for Punjabi dialect and to compare its results with traditional search engine in terms of precision and recall rate curves. The search engine developed will help the users to search and retrieve relevant results in a more efficient and effective manner. This would enhance the productivity and precision for the users of the search engine. With support for duplicate terms, users will now get more relevant results for queries with duplicate terms.

1. The interface of uddan punjabi semantic based search engine.

Figure 4.1: Snapshot of interface of Punjabi search engine.

2. This page shows the results after search process.

3. Results obtained after checking the appropriate synonym word

Figure 4.2: Snapshot of result page

5. ALGORITHM

Step 1: Parse the search string entered into tokens.

Step 2: Store all the tokens in a string array.

Step 3: Initialize the flag status as TRUE.

Step 4: Compare the selected word of the search string with each word of the database of notepad files containing domain information.

Step 5: If the comparison result is TRUE. (a) Change the flag status as FALSE. (b) Match the word with the synonym database.

Step 6: Display the results as SERPS by selecting 10 words preceding the word in the notepad file where match occurs and succeeding 10 words where match occurs store them in array along with the synonym words of the search string.

Step7: Expand the search string at runtime with synonym words from database according to the checked ones and Compare each word of expanded search query with database containing domain information.

Step 8: Display the results as SERPS by selecting 10 words preceding the word in the notepad file where match occurs and succeeding 10 words where match occurs store them in array along with the synonym words of the search string and its meaning from database.

6. RESULTSAND DISCUSSIONS

The traditional search engines they search the web using keyword oriented schemes. It doesn't consider the meanings but just matches the keywords and page is retrieved according to page rank algorithm of the particular search engine. Thus, the results produced are irrelevant and have low accuracy.

Here are some results that have been used for evaluation of this study:

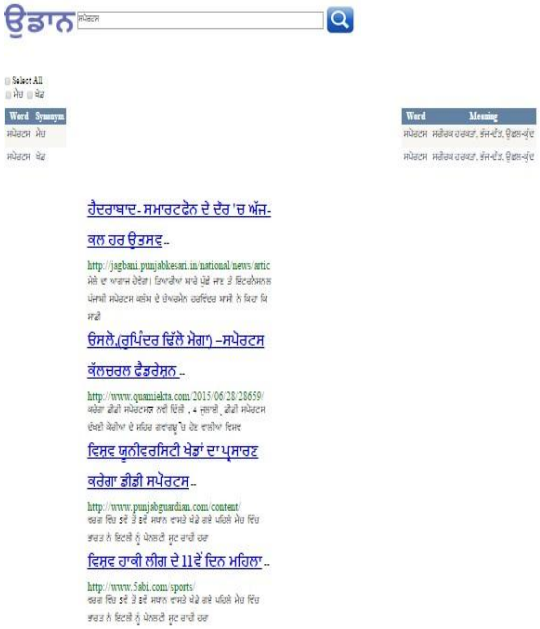

Figure 4.3 Snapshot of result after selecting the synonym word 'Match '

7. ACKNOWLEDGMENTS

I am thankful to my guide Brahmaleen Kaur, Assistant Professor in Department of Computer Science Engineering for their valuable support. Also I am thankful to my Department for the technical support.

8. FUTURE SCOPE

Internet is the largest repository of documents and information. To retrieve relevant results to user query, searching is needed with use of search engine user can locate information quickly from vast amount of knowledge available. The goal of any search is not searching itself but the need of precise and relevant information quickly and with as possible as minimum effort. Through a study of existing research papers it has been found that one of the unexplored languages is Punjabi. Furthermore, searching can be enhanced by developing technique for context based information retrieval.

9. REFERENCES

[1] Wikipedia: http://www.wikipedia.com

[2] N. Stojanovic, "On analysing query ambiguity for query refinement: The librarian agent approach," Lecture notes in computer science, 2003.

[3] "Internet History - Search Engines" (from Search Engine Watch), Universiteit Leiden, Netherlands, September 2001,web:LeidenU-Archie.

[4] Semantic search, Available at: http://www.netlingo.com/lookup.cfm?term=semantic%20search. (Assesed on Nov 2014)

[5] K. S. Esmaili and H. Abolhassani, "A Categorization Scheme for Semantic Web Search Engines," Web Intelligence and Intelligent Agent Technologies, 2009, vol.3, pp. 133-138, 2009.

[6] Senthil J, MargaretAnouncia, andAbhinavKapoor,"Semantic search engine",IJES – Volume 1, Issue 2, November 2013.

[7] Junaidah Mohamed Kassim and Mahathir Rahmany,"Introduction to semantic search engine", International Conference on Electrical Engineering and Informatics, 2009.

[8] QaziMudassarIlyas, Yang Zong Kai and Muhammad AdeelTalib,"A ConceptualArchitecture for Semantic Search Engine", IEEE, 2004.

[9] D. Schneider, D. Stohr, J. Tingvold, A. B.Amundsen, L. Weiland, S. Kopf, W. Effelsberg, A. Scherp," fulgeo—Towards an Intuitive User Interface for a Semantics-enabled Multimedia Search Engine",IEEE, International conference on Semantic Computing,2014

[11] Nandkishor Vasnik, ShriyaSahu, Devshri Roy," Talash: A Semantic and Context Basedoptimized Hindi Search Engine" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[12] Maleerat Sodani and athairat Ketmaneechairat "Information Retrieval Experiment on SubjectiveWords Query Expansion", International Conference of Information and Communication Technology (ICoICT), 2013.

[13] Er.TanjyotAurora, Er.Brahmaleen Kaur, "Asemantic based search engine for Punjabi". International Journal of Information Technology& Computer Sciences Perspectives© Pezzottaite Journals. 1471|Page, Volume 4, Number 2, April – June– 2015 ISSN (Print):2319-9016, (Online):2319-9024 PEZZOTTAITE JOURNALS SJIF (2012): 3.201, SJIF (2013): 5.058, SJIF (2014): 5.891