A Review on Ontology based Information Retrieval System

Mitali Bansal, Jyoti Arora
Research Scholar, Assistant Professor
Department of Computer Science & Engineering, Desh Bhagat University, Mandi Gobindgarh

***Abstract* - Information Retrieval forms an important part of today's world. Extracting information out of unstructured data is an interesting and challenging problem. This paper targets the extraction of important information from house classified data of hindi newspaper. Information Retrieval Vector space model represents the documents and concepts as column and row vector where the cosine angle between the documents are used to measure the similarity between the documents. The similarity between the documents is determined by the closeness of documents in the vector space. This paper attempts to review all works pertaining to Ontology based Information Retrieval System (OBIRS) and their results are discussed according to their performance.**

***Keywords* - Ontology, Information Retrieval System, Domain Knowledge, User Query**

I. INTRODUCTION

Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies. And, being a very active area of research and development, there is not a single agreed-upon definition that would satisfy everyone, but there are some aspects, which would be part of any knowledgeable person's definition. It is thought that humans normally utilize all of these levels since each level conveys different types of meaning. But various NLP systems utilize different levels, or combinations of levels of linguistic analysis, and this is seen in the differences amongst various NLP applications. This also leads to much confusion on the part of non-specialists as to what NLP really is, because a system that uses any subset of these levels of analysis can be said to be an NLP-based system. The difference between them, therefore, may actually be whether the system uses 'weka.'

Ontology based Information Retrieval

Even though search engine technology has experienced impressive enhancements in the last decade, the content description and query processing techniques Information Retrieval (IR) technology currently builds upon are still mostly based on keywords, and therefore provide limited capabilities to capture and exploit the conceptualizations involved in user needs and content meanings. For instance, limitations include the inability to account for relations between search terms (e.g., "hurricanes originated in Mexico" vs. "hurricanes that have affected Mexico", "books about recommender systems" vs. "systems that recommend books"), to handle searches that involve a secondary sense of a term (e.g. "Victor Valdés", the goal keeper vs. "Victor Valdés", the video processing researcher) or to integrate information distributed over several Web resources, (e.g. searches regarding products or services). Aiming to solve the limitations of keyword-based models, the idea of semantic search, understood as searching by meanings rather than literal strings, has been the focus of a wide body of research in the Information Retrieval (IR) and the Semantic Web (SW) communities. However, these two fields have had a different understanding of the problem.

Ontology based web semantics annotation

The World Wide Web is playing a vital role in information sharing for the purpose of business, education, research, etc. A large amount of useful information is available over the web in unstructured, ungrammatical and incoherent formats. This includes reports, scientific papers, product advertisements, news, emails, Wikipedia, etc. In the recent years, it has been seen that the information is also available over the web in various languages such as French, English, Hindi, Arabic, etc. Among variety of languages, the Hindi language is among the largest spoken languages of the world and national language of India. Due to the availability of large amount of Hindi language documents over the web, it is required to develop a sophisticated information retrieval system to utilize the information efficiently. Among these documents there are some online newspapers which are one of the most popular news website in India. The classified is a section of the newspaper where advertisements about different products are published for sell and purchase. These include ads of automobile, real estate, electronic products and so on.

Approaches for ontology based NLP

A number of researchers have attempted to improve the technology for performing various activities that form important parts of NLP work. These activities may be categorized as follows:

- Lexical and morphological analysis, noun phrase generation, word segmentation, and so
- Semantic and discourse analysis, word meaning, and knowledge
- Knowledge-based approaches and tools for NLP

Need for Multilingual approach

Proliferation of multilingual text on the Internet has increased the demand for efficient information retrieval independent of language. Among variety of languages, the Hindi language is one of the most commonly spoken and written language in South Asia. However, due to unstructured format the access of relevant information is still a big challenge. The semantic web technologies enable the advancement in information retrieval systems by assigning semantics to information.

II. RELATED WORK

Fernandez et. al. proposes in his paper that the techniques for content description and query processing in Information Retrieval (IR) are based on keywords, and therefore provide limited capabilities to capture the conceptualizations associated with user needs and contents. Aiming to solve the limitations of keyword-based models, the idea of conceptual search, understood as searching by meanings rather than literal strings, has been the focus of a wide body of research in the IR field. More recently, it has been used as a prototypical scenario (or even envisioned as a potential "killer app") in the Semantic Web (SW) vision, since its emergence in the late nineties. However, current approaches to semantic search developed in the SW area have not yet taken full advantage of the acquired knowledge, accumulated experience, and technological sophistication achieved through several decades of work in the IR field. Starting from this position, this work investigates the definition of an ontology-based IR model, oriented to the exploitation of domain Knowledge Bases to support semantic search capabilities in large document repositories, stressing on the one hand the use of fully fledged ontologies in the semantic-based perspective, and on the other hand the consideration of unstructured content as the target search space.

In a paper, Sonar et. al. present an ontology-based information extraction and retrieval system and its application in the soccer domain. In general, we deal with three issues in semantic search, namely, usability, scalability and retrieval performance. We propose a keyword-based semantic retrieval approach. The performance of the system is improved considerably using domain-specific information extraction, inferencing and rules. Scalability is achieved by adapting a semantic indexing approach and representing the whole world as small independent models. The system is implemented using the state-of-the-art technologies in Semantic Web and its performance is evaluated against traditional systems as well as the query expansion methods. Furthermore, a detailed evaluation is provided to observe the performance gain due to domain-specific information extraction and inferencing. Finally, it is shown how they use semantic indexing to solve simple structural ambiguities.

Uren V. et al. defines an academic document as a document that knows concerning its content in order to progression the information automatically. Classically the information about a document has been organized throughout the use of metadata. Still, the semantic web recommended the annotation of documents use semantic information as of domain ontology's. Semantic annotation officially identifies thoughts and relations among concepts in documents, or it is proposed mainly for used by machines. Furthermore different ontology's presents different conceptualizations; consequently they can be utilize to extract information from the similar document depending lying on the conceptualization individual in the ontology.

Reeve et al. 2010 currently Wimalasuriya et al. and Chi arcos [9] , have offered a comprehensive analysis of semantic annotation tools or categorized them depending on technologies. It have to be pointed that completely automated semantic annotation is at rest an unclear problem due to the verity that annotation involve human intervention in the establishment stage to bootstrap the process. Every existing semantic annotation systems rely on human intervention at number of points, consequently, the annotation process is not completely automated. instruction manual annotation process is not difficult but is an exclusive job in terms of time or cost and therefore becomes unusable for huge size of web contents. since the answer mainly of the reported job was attention at semi automated annotation .

Tao ZangNayak used ontology-based knowledge IR framework which captures user's background knowledge to improve IR performance. Here the ontology is personalized by using the user's local instance repository. The semantic relations of hypernym/hyponym, holonym/meronym and synonym are specified in the ontology model. The performance of the experimental models are measured by three methods: the precision averages at eleven standard recall levels (11SPR), the mean average precision (MAP), and the F1 Measure.

ShlomoBerkovsky · TsviKuflik · Francesco Ricci have worked on similarity-based retrieval of structured data, such as Case-Based Reasoning (CBR). It has been found that in such systems, when the cases contain both numeric and free-textattributes, similarity-based retrieval cannot exploit standard speedup techniques based on multi-dimensional indexing. They have proposed a novel approach for storage of the case-base in a decentralized Peer-to-Peer environment using the notion of Unspecified Ontology to improve the performance of the case retrieval stage and build CBR systems that can scale up to large case-bases. They have developed a distributed an algorithm with distributed nature for the retrieval of approximated mostsimilar cases, which exploits inherent characteristics of the unspecified ontology in order to improve the performance of the case retrieval stage in the CBR problem solving cycle.

Fernandez et. al. invented in his paper the technologies for content explanation or query dispensation in Information Retrieval (IR) are depend on keywords, and after that give incomplete capabilities to capture the generalizations associated by means of user needs or contents. aim to determine the limitations for keyword-based models, the arrangement of abstract search, understand as searching through meanings somewhat literal strings, have been the centered of a extensive deceased of research in the knowledge Retrieval pasture.

In a paper, **Sonar et**. al. present ontology based knowledge extraction or retrieval system or the applications of this into the soccer field. Common we interrelate with three issue into semantic research, scalability, usability and retrieval presentation. We recommend the keyword based semantic retrieval advancement. The presentation of the association is enhanced considerably by using domain specific knowledge extraction, rules or inference. Scalability is achieve with adapt a semantic evidence approach or presenting the complete world as little independent model.

Grauet.al present freshly started EU project 'Optique', which supports for a next generation of the well - known Ontology - Based Data Access (OBDA) approach to address the data access problem in big data. By using Ontology - Based Data Access data is accessed in less time complexity as well as cost for accessing data is reduced. Ontology - Based Data Access strives for query formulation that optimizes the query to manipulate the results in less amount of time. In this paper [9] Author has proposed a general framework for word sense disambiguation by using knowledge latent in Wikipedia.

III. CONCLUSION AND FUTURE SCOPE

A review of various techniques related to ontology based information retrieval system has been discussed. Researchers are utilizing ontology information for improvement in the search relevancy. A novel approach of ontology based information retrieval system has also been discussed which can be applied for classified ads. Various features can be extracted using ontology based rules which has not been dealt in the past to the best of author's knowledge. The results can be compared to prove the effictiveness of the algorithm. In future, other rules can be formed and the designed system can be applied on other databases of other domains. Also it can be formulated for other foreign languages and a hybrid system can be developed.

IV. REFERENCES

[1] Kara, Soner, et al. "An ontology-based retrieval system using semantic indexing." *Information Systems* 37.4 (2012): 294-305.
[2] Fernández, Miriam, et al. "Semantically enhanced Information Retrieval: an ontology-based approach." *Web Semantics: Science, Services and Agents on the World Wide Web* 9.4 (2011): 434-452.
[3] Ahmad Z., Khan M A, Ali R, Ahmad I, Amir M. Evolving Web Corpus: Text Powered by Non Text. Conference on Language and Technology (CLT), Islamabad, Pakistan, 2010.
[4] Rajput Q, Haider S. BNOSA: A Bayesian Network and Ontology Based Semantic Annotation Framework. Journal of Web Semantics Science, Services and Agents on the World Wide Web, 9(2), 2011, pp. 99-112.
[5] Antoniou G, Harmelen F V. A Semantic Web Primer.MIT Press, 2007.
[6] Uren V, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, Motta E, Ciravegna F. Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 4(1), 2006, pp. 14–28.
[7] Lee T B. The Semantic Web, Scientific American. 284 (5), 2001, pp. 34-43.
[8] Abbas Q. Building a Hierarchical Annotated Corpus of Hindi: The HINDI.KON-TB Treebank. Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg,2012, pp. 66–79.
[9] Rajput Q, Haider S. A Comparison of Two Ontology-based Semantic Annotation Frameworks. Artificial intelligence application and innovations (IFIP), vol. 339, 2010, pp. 187-194.
[10] Rajput Q, Haider S. A Comparison of Ontology-based and Reference-Set-based Semantic Annotation Frameworks. Proceedia Computer Science, vol. 3, 2011, pp. 1535-1540.