A Novel OBIRS System for 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. A Novel Ontology based Information Retrieval System (OBIRS) is developed in this thesis and the results are obtained according to user query. The implementation is done using Python and GUI is developed using tkinter.

I. Introduction

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. Jan Paralic and Ivan Kostial [1] developed an Ontology based information retrieval, where the user query is preprocessed using Ontology to retrieve the set of concepts, which is compared with the concepts in the document vector. Weights are drawn for tho se of individual concepts in the document vector using which relevant documents are retrieved. Pablo Castells et al. [2] proposed semantic based personalization for Ontology based information retrieval and an adaptation of Ontology in information retrieval. Personalization method works by triggering a formal Ontology based query followed by vector space information retrieval technique. The cosine similarity of concepts and keywords with documents as vector is used.

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, reviews, 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 Chinese, French, English, Hindi, Hindi, Arabic, etc. 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

The data collection stage is critical to all three approaches although statistical and connectionist approaches typically require much more data than symbolic approaches. In the data analysis/model building stage, symbolic approaches rely on human analysis of the data in order to form a theory while statistical approaches manually define a statistical model that is an approximate generalization of the collected data.

This paper proposes a novel Ontology based Information Retrieval System (OBIRS) for extracting meaningful information out of hindi classified data.

Rest of the paper is as follows. . Section II describes the literature survey taken from various papers which have discussed relevant works. Section III discusses our problem statement and Section IV provides a description of our proposed methodology and results are put together in section V. Finally the paper concludes with a brief discussion the overall results.

II. LITERATURE SURVEY

Recommender systems have become an important research area since the emergence of the first research paper on collaborative filtering in the mid-1990s [4]. In general, recommender systems directly help users to find content, products, or services (such as books, digital products, movies, music, TV programs, and web sites) by aggregating and analyzing suggestions from other users, which mean reviews from various authorities, and users. Recommender systems are generally classified into collaborative filtering (CF) and content-based filtering (CB) [5]. In general, CF uses an information filtering technique based on the user's previous evaluation of items or history of previous purchases. However, this technique has been known to reveal two major issues: sparsity problem and the scalability problem In contrast, CB analyzes a set of documents rated by an individual user and uses the contents of the documents, as well as the provided ratings, to infer a user profile that can be used to recommend additional items of interest [6].

Being an active and comparatively new area of research, a lot of techniques for semantic annotation have been proposed. Reeve et al. 10 and recently Wimalasuriya et al. and Chiarcos, have presented a comprehensive survey of semantic annotation tools and categorized them depending on the methodologies. It must be noted that fully automated semantic annotation is still an unsolved problem due to the fact that annotation requires human intervention in the beginning stage to bootstrap the process. All existing

semantic annotation systems rely on human intervention at some points, therefore, the annotation process is still not completely automated.

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.

III. Problem Statement

This thesis presents a semantic annotation framework that can annotate documents written in Hindi language. The framework uses domain specific ontology and context keywords instead of NLP (Natural Language processing) techniques. The experiment has been conducted to evaluate the presented annotation framework. The set of corpora used in the experiment belong to the online classified ads posted on the online Hindi newspapers.

The presented semantic annotation framework comprises on the following two components:

- Construction of domain ontology and
- Information extraction using ontology for annotation

Objectives

The major objectives of this thesis can be summarised as follows:

- Creation of house data for hindi classified newspapers
- Creation of rules for web semantics related to Hindi language
- Pre-processing of the documents
- Creation of ontologies for that particular domain
- Application of natural language processing and ontologies like context keyword, format and additional knowledge for extracting information out of classified ads.
- Creation of database in a structured query language
- Retrieval of the data in a structured manner

IV. Proposed Methodology

The thesis aims at application of ontology based web semantics annotation using natural language processing. The ontology is constructed manually by conceptualizing the domain knowledge. The domain knowledge consists of concept and properties of a domain that are generally searched by user and thus specified in ontology. An example of our dataset for house ad is given as follows:

२ बी एच के म्कान किराए पे लगाने के लिए सबसे सस्ते दाम मे| १३०० वर्ग फुट मे बना यह म्कान आलीशान है और इसमे दो फ्रिड्ज एक टीवी और हर कमरे मे एसी लगा हुआ है| किराया २०००० रु प्रति माह होगा और एक महीने का अड्वान्स जमा करना होगा| यह नीरवाना सोसाइटी मे सेक्टर ४९, चंडीगढ़ मे स्थित है|

The keywords are: rent, area, type of house, etc.

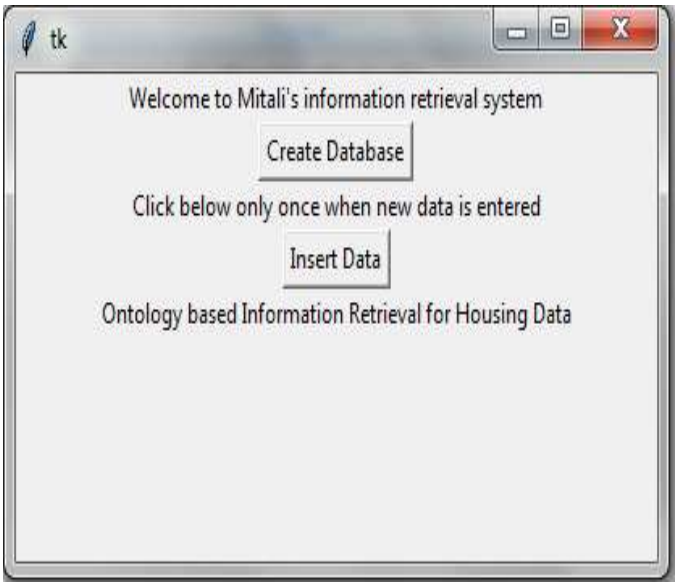
Once information is specified in the ontology, the extraction step generates rules for each data-type used in ontology to extract values of each property. Rules are generated as regular expression according to the corresponding data-type as specified in the ontology to represent the lexical appearance of data

The information extraction algorithm searches the context keywords in the document, if a match is found, this suggests that the corresponding value will be found in the neighborhood of this keyword and can be extracted by using corresponding rules. The extraction process will continue for each property as specified in ontology. Once all the data is extracted, it is stored in the database where structured query can be applied.

Rules are formed using ontology in house ads domain and the data is crawled for various information. The information is organized and stored to specific keywords which are given below:

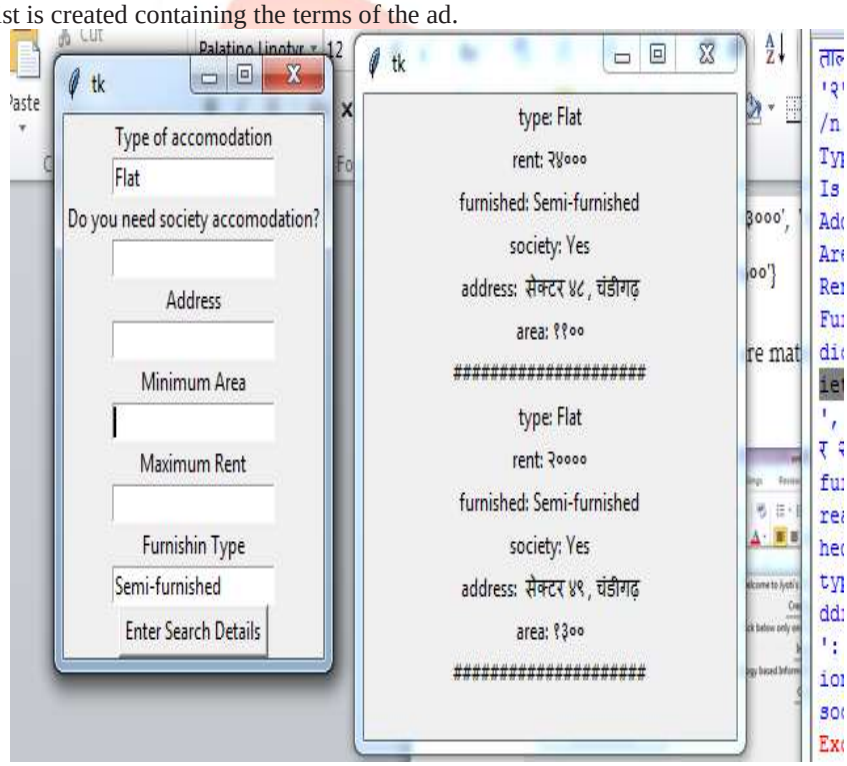
1. Type of accommodation
2. Whether a society accommodation
3. Rent
4. Area
5. Furnishing
6. Address

V. Results and Discussion



The second part is created for entering the classified ads into the database. A button named as Insert Data is made which should only be pressed once whenever there are new entries. Pressing this button more than once will create duplicate copies of entries which will slow down the search process during information retrieval.

The major part of this system is the information retrieval system. This system is used for extracting the information out of the raw ad data. For each data a list is created containing the terms of the ad.



The results are found to be quite accurate in most of the cases and hence it proves the efficacy of our algorithm.

VI. Conclusion

A novel approach of ontology based information retrieval system is designed for classified ads. The ads database was taken for house data which are given in hindi newspaper. Various features were extracted using ontology based rules which has not been dealt in the past to the best of author's knowledge. The results were found to be quite encouraging and proves 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.

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] Syed A Z, Aslam M, Martinez-Enriquez A M. Lexicon Based Sentiment Analysis of Hindi Text Using SentiUnits. Advances in Artificial Intelligence, Springer Berlin Heidelberg, 2010, pp. 32–43.