Information Retrieval Techniques based on Ontology for High Effectiveness

Komal Shivaji Mule
M.E Scholar
Dr. D.Y.Patil School Of Engineering & Technology, Lohegaon, Pune
Affiliated to Savitribai Phule Pune University

Arti Waghmare
Assistant Professor
Dr. D.Y.Patil School Of Engineering & Technology, Lohegaon, Pune
Affiliated to Savitribai Phule Pune University

ABSTRACT
Information Retrieving is task of recuperating information with high relevance, precision and recall. Basic methods for information retrieval include Boolean Retrieval, Fuzzy retrieval, Vector Space model. Searching depends on matching keywords between user-query and document. Ontology can be used in information retrieval. In software engineering and information science, ontology is a formal naming and meaning of the types, properties, and interrelationships of the elements that truly or in a broad sense exist for a specific domain of discourse. Ontology provides us with vocabulary of terms. New terms and relations could be found out using the existing relations. Many Information retrieval techniques exist amongst which many depend on keywords. Relations between keywords are considered here so as to get higher precision and recall.

General Terms
Internet Searching, Spidering, Information Separating

Keywords
Information Retrieval (IR), Ontology

1. INTRODUCTION
Information retrieval can span many areas of computer science. IR can be used areas such as media search, search engines, etc. domain specific application of IR are expert search finding, geographic IR, IR for chemical structure, IR in software engineering, Legal IR and vertical search. Information can be retrieved through huge collection of database. The relevance and effectiveness of data that is retrieved depend upon the arrangement structure of database. The utilization of metadata will become important with the increase in World Wide Web. IR & information filtering are diverse function. IR is expected to support individuals who are effectively looking or hunting for data as in internet searching. IR typically assumes static database or related static database against which individual seek information. Web search technology built this database by sending out spiders and afterwards indexing web pages that are discovered. Information separating supports to find people desired information. Boolean model refers to keywords combined using AND, OR, NOT. AND refers to intersection OR refers to union. In OR query result is retrieved even if one term is present in the database. AND query retrieves result if and only if all the terms are present in static database. No ranking is there in Boolean model. Vector space model or term vector model is an arithmetical model for representing content documents as vectors of identifiers, for instance, index terms. It is utilized as a part of data separating, data recovery, indexing and pertinence rankings.

1.1 Word about Ontology
In software engineering and computer science, ontology is a formal naming and meaning of the types, properties, and interrelationships of the substances that truly or in a far-reaching way exist for a specific area of talk. It is in this manner a viable utilization of philosophical ontology, with scientific classification i.e taxonomy. Ontology compartmentalizes the variables required for some set of computations and creates the connections between them. Ontology is triplet of subject, predicate, object which is mathematically represented as (S, P, O) where S represented for subject, P stands for predicate and O stands for object. Consider an example where ‘john has employer at ACC Inc.’. This example shows that ‘john’ is subject, ‘ACC Inc’ is object and ‘has employer’ is predicate i.e. the relation between ‘john’ and ‘ACC Inc’. Figure 1 shows the example of ontology.

2. LITERATURE SURVEY
In the previous few years, there has been an exponential increment in the measure of data accessible on the World Wide Web. This plenty of data can be greatly beneficial for clients. However, the measure of human mediation that is as of now needed for this is inconvenient. Information extraction (IE) frameworks attempt to take care of this issue by making the assignment as programmed as could be possible. Most of the systems require user feedback during the information extraction. Fatima et. al. proposed a framework that utilizes grouping systems for automatic IE from HTML records containing semi structured information. Utilizing domain specific data gave by the client; the proposed framework parses and tokenizes the information from a HTML archive, segments it into groups containing comparable components, and quotes an extraction guideline in light of the example of event of data tokens. The extraction guideline is then used to refine clusters, and finally, the output is reported. A multiobjective genetic-algorithm-based clustering approach is utilized; it is fit for finding the quantity of clusters and the most regular clustering [5].

An expanding number of databases have ended up web open through HTML structure based search interfaces. The data units came back from the fundamental database are typically encoded into the outcome pages powerfully for human surf. For the encoded information units to be machine 6, which is crucial for some applications, for example, profound web information gathering and Internet comparison shopping, they have to be extricated out and relegated significant names. Author proposed an annotation approach that first adjusts the information units on an outcome page into distinctive gatherings such that the information in the same gathering have the same semantic. At that point, for every gathering

author explain it from distinctive perspectives and total the diverse annotations to foresee a last annotation mark for it. An annotation wrapper for the pursuit webpage is consequently built and can be utilized to explain new result pages from the same web database [6].

Jiang et. al. [7] presents Forum Crawler Under Supervision (FoCUS), a supervised web-scale forum crawler. The objective of FoCUS is to crawl applicable forum content from the web with negligible overhead. Forum threads contain data content that is the focus of gathering crawlers. In spite of the fact that discussions have distinctive formats or styles and are controlled by diverse discussion programming clusters, they generally have comparable implied route ways joined by particular URL sorts to lead clients from entrance pages to string pages. In view of this perception, Author decreased the web discussion slithering issue to a URL-sort distinguishment issue. Also, Author demonstrates to learn precise and viable customary statement examples of understood route ways from consequently made preparing sets utilizing accumulated results from feeble page sort classifiers. Powerful page sort classifiers can be prepared from as few as five commented discussions and connected to a substantial arrangement of concealed gatherings. Test outcomes demonstrate that FoCUS accomplished more than 98 percent adequacy and 97 percent scope on an expansive arrangement of test discussions fueled by more than 150 distinctive discussion programming clusters.

It is decently perceived that the Internet has turned into the biggest commercial center on the planet, and web publicizing is extremely mainstream with various businesses, including the customary mining administration industry where mining administration commercials are viable transporters of mining administration data. Nonetheless, benefit clients may experience three noteworthy issues – heterogeneity, omnipresence, what's more, uncertainty, when hunting down mining administration data over the Internet. In this paper, author display the structure of a novel self-versatile semantic centered crawler – SASF crawler, with the reason for exactly and efficiently finding, organizing, what's more, indexing mining administration data over the Internet, by considering the three noteworthy issues. This structure consolidates the advances of semantic centered creeping and philosophy adapting, with a specific end goal to keep up the execution of this crawler, paying little respect to the assortment in the Web environment. The advancements of this exploration lie in the configuration of an unsupervised structure for vocabulary-based philosophy learning, and a mixture calculation for coordinating semantically important ideas and metadata. A progression of investigations is directed with a specific end goal to assess the execution of this crawler [8].

Reference paper [9] uses ontology for Information Retrieval. Deficiency reliance (D)-grid is an orderly demonstrative model to catch the progressive framework level flaw analytic data comprising of conditions between noticeable indications and disappointment modes connected with a framework. Building a D-grid from first standards and overhauling it utilizing the space learning is a work concentrated and prolonged undertaking. Further, in-time enlargement of D-lattice through the disclosure of new indications and disappointment modes watched for the first time is a testing assignment. Here, author portrays a metaphysics based content digging strategy for consequently building and upgrading a D-grid by mining countless repair verbatim (regularly written in unstructured content) gathered amid the conclusion scenes. In this methodology, author has first build the issue conclusion cosmology comprising of ideas and connections regularly saw in the flaw analysis space. Next, they utilize the content mining calculations that make utilization of this metaphysics to distinguish the vital relics, for example, parts, indications, disappointment modes, and their conditions from the unstructured repair verbatim content.

3. USE OF ONTOLOGY FOR IR

Various techniques have been developed for information retrieval. Most of IRS is based on keywords. Problem of keywords is mentioned in [11][12]. Relation between the keywords and phrases are not mentioned in these systems due to which precision and recall rates are low. The mapping of concepts in information into conceptual models, i.e. ontologies, gives off an impression of being a valuable technique for moving from keyword based to concept based data retrieval. Frequency of keywords can be calculated. Distance between keywords can be calculated.

Use of ontology yields high relevancy for the information retrieval. Jan and Ivan in [10] have described the ontological method for information retrieval. Domain knowledge is represented in the form of ontology. Webocrat system is used here. The Webocrat system has been developed within the EC funded project IST-1999-20364 Web Technologies Supporting Direct Participation in Democratic Processes (WEBOCRACY). It is an ontology based model meant for providing quality services to citizens. Domain knowledge is linked with Webocrat system. So here exact match is not made unlike the vector model. Match is made here depending on the relation in the ontological model. Figure 2 describes the information retrieval model designed by Jan and Ivan.

Reference [1] describes the ontological indexing. Languages used for representing ontology are OWL, RDF, and OIL. RDF facilitates encoding, exchange and reuse of structured metadata. Objectives and goals of RDFs are independence, scalability, interchange. Firstly the ontologies are imported and mapped. For this firstly the RDF-file is converted to JENA model. Secondly configuration file for ontology import is read. Finally concepts and relations are written into the database. Spidering and indexing is done to create the domain ontology. Spidering and indexing includes Spidering across web and storing the data into database after removal of stop words and bad words. Ontological descriptors are situated and generative ontology is created with the noun phrases. Zones are created to store data of similar data type. Zones are nothing but clusters with similar data stored in a particular zone. Inverted indices are used here. Inverted index is a an index data structure putting away a mapping from substance, for example, words or numbers, to its areas in a database record, or in an document or an arrangement of documents. Tag based rank algorithm is used here. The algorithm is as follows:

i. Tag based rank is calculated using

$$tbr = \sum \text{weighting } (w_i, t_i)$$

ii. Drop XML Tags

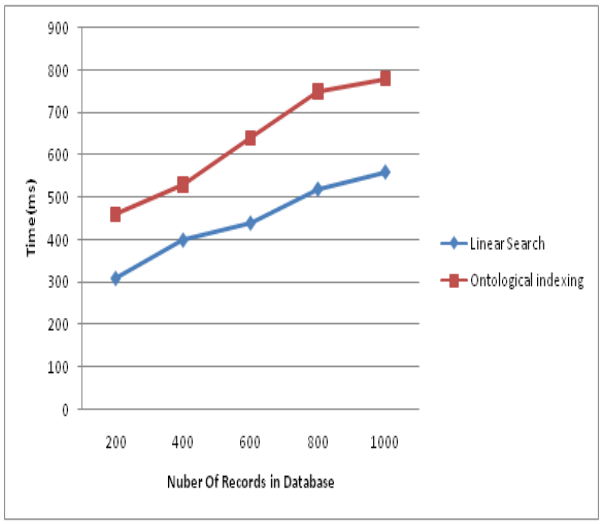
iii. Words and their ranking is written to zone indexing in database

Figure 3 explains that user fires query to search engine. Search engine searches for similar words in inverted indices and also searches for its synonyms. WordNet is used here for this purpose.

A separate dictionary can be built for the words that are not in WordNet dictionary. In case exact match is not found then suffixes are constructed from the keywords for getting the match.

4. RESULTS

Comparison of ontological indexing is done with linear search experimentally. Here, x-axis is number of records in database and y-axis denotes time taken to retrieve result in milliseconds.



Graph 1: comparison of ontological indexing with linear search

5. FIGURES

Fig 1: Example of Ontology

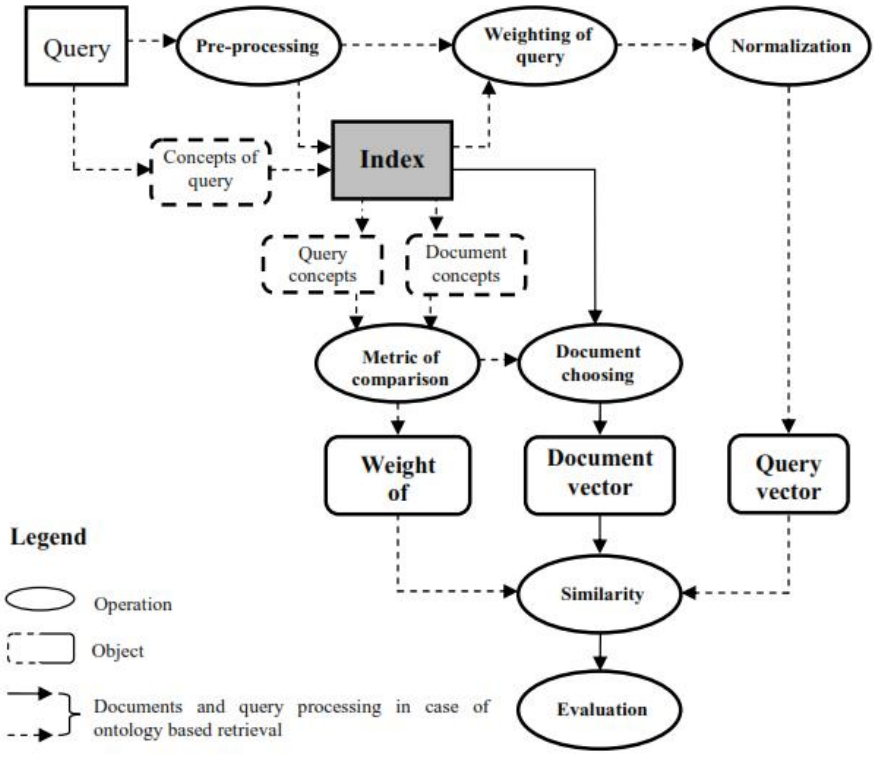


Fig 2: Ontology based information retrieval

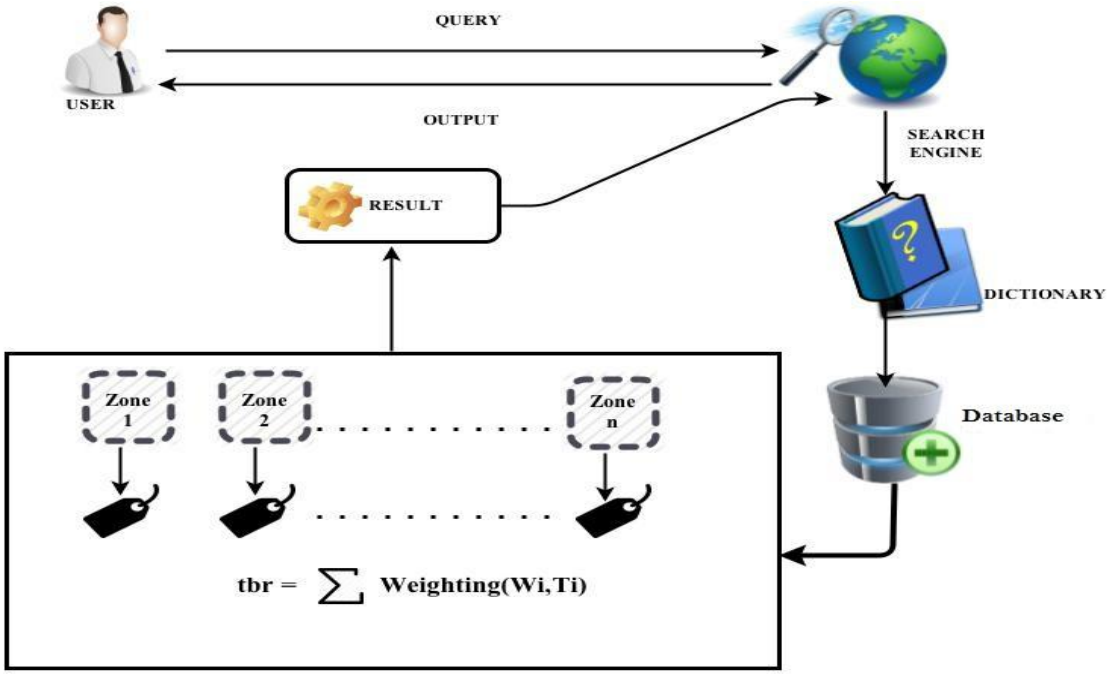


Fig 3: Architecture Diagram

6. CONCLUSION

Ontology helps retrieve information effectively with high precision and recall. Ontology facilitates to find relation between keywords. The relevance of results is high with use of ontology for Information Retrieval. Inverted indices are used for indexing. In future more focus will be given on the keywords.

7. ACKNOWLEDGMENTS

I would like to thank the researchers as well as publishers for making their resources available.

8. REFERENCES

[1] Rajeswari Mukesh, Sathish Kumar Penchala, and Anupama K. Ingale. Ontology Based Zone Indexing Using Information Retrieval Systems. S. Unnikrishnan, S. Surve, and D. Bhoir (Eds.): ICAC3 2013, CCIS 361, pp. 181–186, 2013.

[2] TroelsAndreasen, HenrikBulskov. Conceptual querying through ontologies. 2009 Elsevier.

[3] Eduard Dragut, Fang Fang, Prasad Sistla, Clement Yu. Stop Word and Related Problems in Web Interface Integration. August 24-28, 2009.

[4] Saruladha, K., Aghila, G., Penchala, S.K., "Design of New Indexing Techniques Based on Ontology for Information Retrieval Systems", In: Das, V.V., Vijaykumar, R. (eds.) ICT 2010. CCIS, vol. 101, pp. 287291. Springer, Heidelberg (2010)

[5] Fatima Ashraf, Tansel Ozyer, And Reda Alhajj,Associate Member, IEEE,"Employing Clustering Techniques For Automatic Information Extraction From Html Documents",IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 38, No. 5, September 2008

[6] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Member, IEEE, and Clement Yu, Senior Member, IEEE,"Annotating Search Results from Web Databases", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 3, March 2013

[7] Jingtian Jiang, Xinying Song, Nenghai Yu, Member, IEEE, and Chin-Yew Lin, Member, IEEE, "FoCUS: Learning to Crawl Web Forums", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 6, June 2013

[8] Hai Dong, Member, IEEE, and Farookh Khadeer Hussain, "Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery", IEEE Transactions On Industrial Informatics, Vol. 10, No. 2, May 2014

[9] Dnyanesh G. Rajpathak, Member, IEEE and Satnam Singh, Senior Member, IEEE, "An Ontology-Based Text Mining Method to Develop D-Matrix from Unstructured Text", IEEE Transactions On Systems, Man, And Cybernetics: Systems, Vol. 44, No. 7, July 2014

[10] Jan Paralic, Ivan Kostial, "A Document Retrieval Method Based On Ontology Associations", Original Scientific Paper

[11] O. Dridi, RIADI Laboratory, National School of Computer Sciences, Tunisia, "Ontology-Based Information Retrieval :Overview and New Proposition"

[12] Jinwoo Kim, Dennis McLeod, "A 3-Tuple Information Retrieval Query Interface with Ontology Based Ranking", IEEE IRI 2012