Improved Indexing Technique for Information Retrieval based on Ontological Concepts

Komal Shivaji Mule
M.E Scholar,
Dept. Of Computer Engg.,
Dr. D. Y. Patil School OfEngg& Tech,
Affiliated to SavitribaiPhule Pune University, India,

Arti Waghmare
Assistant Professor,
Dept. Of Computer Engg.,
Dr. D. Y. Patil School OfEngg& Tech,
Affiliated to SavitribaiPhule Pune University, India,

ABSTRACT

Ontology has a richer internal structure as it includes relations and constraints between the concepts. Ontology can be used for information retrieval. Ontology is a halfway determination of a conceptual vocabulary to be utilized for formulating knowledge-level hypotheses around a domain of discourse. The key part of ontology is to help knowledge sharing and reuse. The process of allotting descriptions to documents in an IRS is called indexing. In previous system zone based indexing is introduced which has certain drawbacks. It helps finding results of user's query with exact match. A new technique is proposed which improves results. In this technique web pages are stored in xml database. Zones are formed in database. In case exact match is not found in xml database using zone based indexing then proximity of keyword is retrieved from the n-ary tree which is constructed using ontology. WordNet is used as dictionary for finding related words similar to user's query. A separate dictionary is created for words that are not present in WordNet. This application can be implemented in Libraries for access of books. Even if exact match is not available then also some of the related books can be retrieved. The aim of proposed system is to achieve higher Retrieval Status Value.

Keywords

Information Retrieval,ontology, RSV, n-ary, Zone based indexing.

1. INTRODUCTION

1.1. Information Retrieval

Information retrieval is the task, of discovering the important records, given a set of reports and a client inquiry. Information retrieval is a field of study that helps the user to find needed information from a large collection of text documents. Information retrieval applications oblige speed, consistency, exactness and convenience in recovering significant writings to fulfill client inquiries. The fast development of the Web in the most recent decade makes it the biggest freely available information source on the planet. Searching the Web is increasingly becoming the leading information seeking method due to web search convenience and the richness of information on the Web. People make fewer and fewer trips to libraries, but more and more searches on the Web. In information retrieval the roots of web search are situated. IR tools that index the Web give an expansiveness and straightforward entry to data that was mindboggling just 10 years back. IR has additionally become vital at the other end of the size spectrum. For instance, the help services incorporated with OS depend on efficient content search, and search systems help clients to find files on their personal computers. Figure-1 gives general IR system architecture.

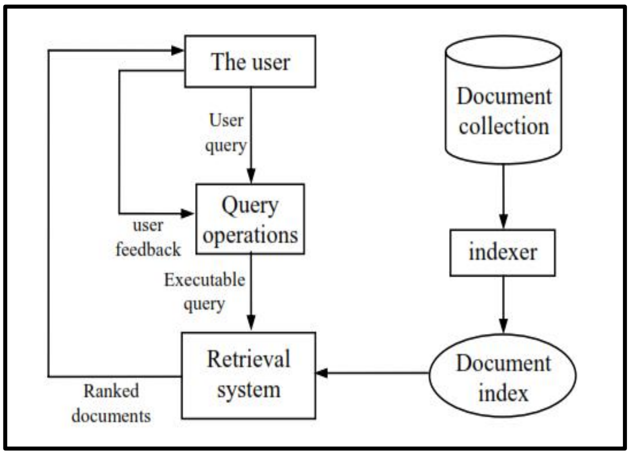

Figure-1: Information Retrieval system architecture

IR systems were initially created to help oversee the vast scientific literature which was created subsequent to the 1940s. Numerous colleges, corporate, and open libraries now utilize IR frameworks to give access to books, diaries, and different records. IR has been discovered valuable in such different territories as office computerization and programming building.IR is utilized today as a part of numerous applications. It is utilized to hunt down records, substance thereof, and record metadata inside customary social databases or web reports. Figure-2 below illustrates the IR system.

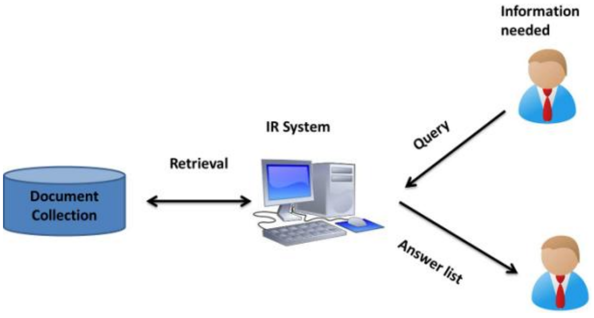

Figure-2: IR system

1.1.1. IR Algorithms

We classify IR algorithms into three categories i.e Retrieval Algorithms, filtering algorithm, indexing algorithm.

a) The Retrieval algorithm

Deals with extracting information from textual database.
Retrieval algorithms are of two types:

- Sequential scan: It involves sequential scanning of textual database
- Indexed text: an "index" of the text is available, and can be used to speed up the search. The index size is normally proportional to the database size, and the search time is sub linear on the size of the text, for instance, signature files and inverted files.

b) Filtering Algorithm

This class of Algorithms is such that the content is the input and a transformed or sifted variant of the content is the output. This is a commonplace transformation in IR, for instance to diminish the size of a message, and/or standardize it to rearrange searching.

The most common filtering operations are:

- Common words removed using a list of stop words;
- Uppercase letters transmute to lowercase letters;
- sequences of multiple spaces reduced to one space and Special symbols removed;
- Numbers and dates transmuted to a standard format;
- Word ranking.
- Word stemming (removing suffixes and/or prefixes);
- Automatic keyword extraction;

1.1.2. Concept of Ontology

Ontology is a set of axioms. Axioms are expressed in an ontology language. Ontology is an explicit specification of a conceptualization. The term is obtained from philosophy, where ontology is an efficient record of Existence. For knowledge-based frameworks, what "exists" is precisely that which can be represented [11]. Ontology is a triplet of subject, predicate and object, where predicate associates the subject and object. For example, 'mother' and 'child' are subject and object respectively and 'parent of' is predicate. So it can be represented as 'mother is parent of child'. Diagrammatically it can be represented as follows.

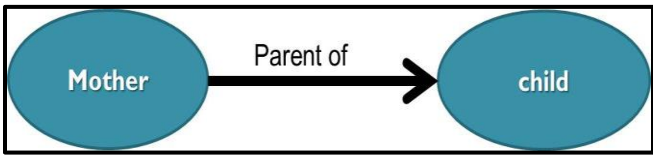

Figure-3: Example of ontology

Subject and object are represented by oval shape and predicate is represented by unidirectional arrow. Ontology is a halfway determination of a conceptual vocabulary to be utilized for formulating knowledge-level hypotheses around a domain of discourse. The key part of ontology is to help knowledge sharing and reuse.

2. RELATED WORK

Conceptual querying is introduced in [3], where concepts are mapped through ontologies instead of dealing with whole documents. Concepts are extracted through the documents and mapped through ontology. Tremendous measure of outsource data are stored and retrieved over the worldwide [4]. Amid that process some troubles are raised to keep up security while giving retrieval and searching methodology. Because of security concerns, delicate information is ensured by encryption before moving to the cloud. Regularly exact data recovery is troublesome over encoded cloud information. To conquer these issues, author proposed an Upgraded Multikeyword Top-k Search and Retrieval (EMTR) scheme; it accomplishes great exactness and high proficiency. To start with, the record can be productively assessed utilizing inverted indexing. At that point the client could query by any number of questioned essential words showing up in the record which assess the pertinence scoring of the record and to the pursuit question to recover applicable data from distributed storage. Another positioning methodology is proposed to recover most noteworthy positioned archives (i.e., most significant) in the information set that brings extensive speedup over inverted weighted indexing.

The Semantic Web acknowledgment is focused around the accessibility of a noteworthy mass of metadata for the web substance, connected with the fitting data about the world. Looking the unlimited web for little specific data prompts numerous anomalies in the consequences of present searching techniques, for example, imprecision, extensive output, not able to interpret the feeling of client's query etc. With the reason for defeating the pitfalls of the current methodologies, author has proposed construction modeling of semantic data recovery to upgrade the pertinence of indexed lists. An algorithm is proposed to process the rank of the reports and afterward semantic indexing is being performed by these positioned website pages. In the algorithm, author is considering the two variables for computing the page rank; one is the recurrence of the essential word happened in the website page and an alternate is affiliated component of the same catchphrase with the important interrogative words which are by and large overlooked by the current pursuit plans. Author has tried algorithm for different records and thus the proposed methodology gives the most pertinent results on the highest point of the result set [5].

Data Retrieval manages recovering archives from an extensive accumulation that matches the data need of a client. Proficient retrieval is focused around the correct stockpiling of the inverted index. There have been numerous systems for diminishing the extent of the inverted index. Static index pruning is one such strategy, which is utilized to diminish the index size. Author investigates a static index pruning methodology which is helpful to lessen the file size. The proposed methodology prunes the whole document from the index focused around its vitality and pertinence of top-k results. The end happens on the premise of the score of the individual report. Investigations have been directed on the FIRE content gathering. In light of the results, it was observed that for particular accumulations, the proposed model gives better accuracy values for the recovery of main 30 or more reports.[6]

To make machine comprehend the semantics of site page, there is requirement for representation dialect other than HTML, XML. Semantic Web permits the data to be represented in decently characterized way utilizing diverse dialects like RDF, OWL which improves the derivation force of the machines, making the substance machine-interpretable. Indexing of the fetched web substance for these representation dialects for viable information retrieval is the key issue of research.

In reference [7] investigation of existing indexing strategies for RDF archives is carried out. A framework is proposed and implemented crawled indexing ontologies represented in semantic web language like RDF.

Reference [8] shows the use of ontology in information retrieval. To enhance the significance of information retrieval, author proposed a methodology focused around the utilization of domain ontology for indexing a collection of documents and the utilization of semantic links between records in the collection to permit the induction of all significant documents. The work includes the execution of a framework focused around the utilization of OWL ontology for pedagogical reports. For this situation, the descriptors are not specifically picked in the documents however in the ontology and are listed by concepts that reflect their meaning instead of words. To perform a search focused around meaning, documents and their descriptors are put away in OWL ontologies depicting the narrative peculiarities of a document. The target is to plan two sorts of OWL ontologies: archive ontology saved for storage of all pedagogical documents and domain ontology held for decently organized of documents stored away in the level of the document ontology and each document is indexed by its keywords and their synonyms. Reference [10] discusses about various ontology based techniques for information retrieval.

Scent of information and content of the clicked URLs is used in [12] for improving the precision of the information retrieval. Information available in geographically distributed area is sometimes knotty to retrieve [13].Grid based Information retrieval is equipped for seeking endless measures of geographically circulated data, which couldn’t be sought viably, if by any means, by traditional information retrieval.

3. PROPOSED WORK

3.1. Problem Definition

User submits query with the hope of retrieving set of relevant documents as result. In existing system exact match is searched in user’s query. In case exact match is not found no results are retrieved. Users cannot be totally reliable on the retrieved results. In case exact results are not found users should get at least related results.

3.2. Proposed system

Zone based indexing with n-ary tree is proposed. Zones are nothing but clusters with similar concepts. Zones are formed by using xml database.

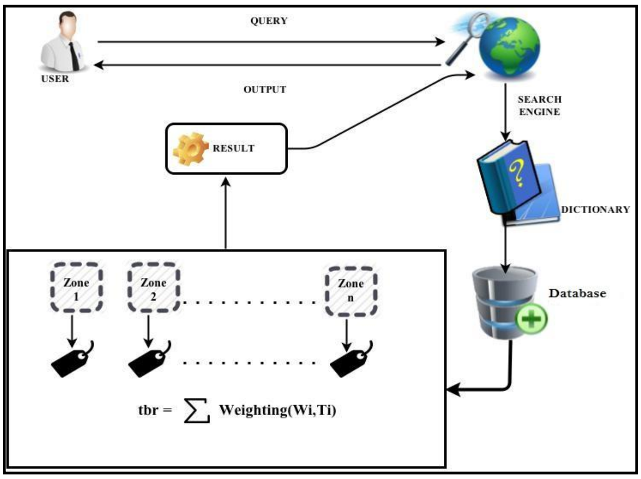

Figure-4: System Architecture of proposed model

As given in figure-4

1. Firstly user queries the search engine.
2. Stopwords are removed from the query
3. The search engine checks for words with similar meaning in WordNet. A separate dictionary is built for words not present in WordNet.
4. The keywords that match with zones retrieve the query to user.
5. In zone based indexing exact words are matched. In case exact result is not found the concept of n-ary tree is integrated with zone based indexing.

For instance if we query the search engine “Sachin is a good batsman”. Then the first word ‘Sachin’ will be searched in database. i.e firstly the letter ‘s’ will be searched. If match is found in database then next letter ‘a’ will be compared. If match found then next letter ‘c’ will be considered for comparison. If the word ‘sachin’ is found then space is there in query. Space is called as tag and is removed from query. Then we come across word ‘is’ and ‘a’ which are stop words, and are removed from query. A list of stop wordsand bad wordsare stored in database. Bad words are vulgar words that come across web. In this technique [1] result is retrieved in case of exact match of words. In case exact match is not found n-ary tree is used to give results that are nearer to the query [9]. Data across web pages is stored in xml database. Ontological concepts are adjectives or nouns that come across database.

4. MATHEMATICAL MODELLING

Each document or query is labeled as a “bag” of words or terms. That is, a document is described by a set of distinctive terms. A term is simply a word whose semantics helps remember the document’s main themes. Each term is associated with a weight. Given a collection of documents D, let V = $\{t_1, t_{2,} \ldots ,t_{|v|}\}$ be the set of distinctive terms in the collection, where t_i is a term. The set V is usually called the vocabulary of the collection, and |V| is its size, i.e., the number of terms in |V|. A weight $W_{ij}> 0$ is associated with each term t_i of document $d_j\in$ D. For a term that does not appear in document $d_{j,}$ $W_{ij} = 0$. Each document d_j is thus represented with a term vector, $_{dj=}$ $\{W_{1j}, W_{2j}, \ldots,$

$W_{|v|j}$},where each weight W_{ij} corresponds to the term $t_i \in V$ and quantifies the level of importance of t_i in document d_j.

Following is the mathematical representation of ontology. A domain ontology is a formal representation of a domain knowledge, by a set of concepts $\mathcal{C}$ within the domain and the relationship $\mathcal{R}$ among them. An ontology $\mathcal{O}$ is defined as structure $\mathcal{O} := (\mathcal{C}, \leq_{\mathcal{C}}, \sigma, \mathcal{R}, \leq_{\mathcal{R}}, \mathcal{A})$consisting of:

- Two disjoint set of concepts/classes, which are entities in the ontology domain, and $\mathcal{R}$ is a set of relations defined among the concepts,
- A partial order $\leq_{\mathcal{C}}$ $\mathcal{C}$called concept hierarchy or taxonomy.
- A function $\sigma : \mathcal{R} \rightarrow \mathcal{C}^{+}$ called signature (where $\mathcal{C}^{+}$ is the set of all finite tuples of elements in $\mathcal{C}$),
- A partial order $\leq_{\mathcal{R}}$ on $\mathcal{R}$, called relation hierarchy, and

A set $\mathcal{A}$ of logical axioms in some logical language $\mathcal{L}$, that can describe constraints on the ontology.

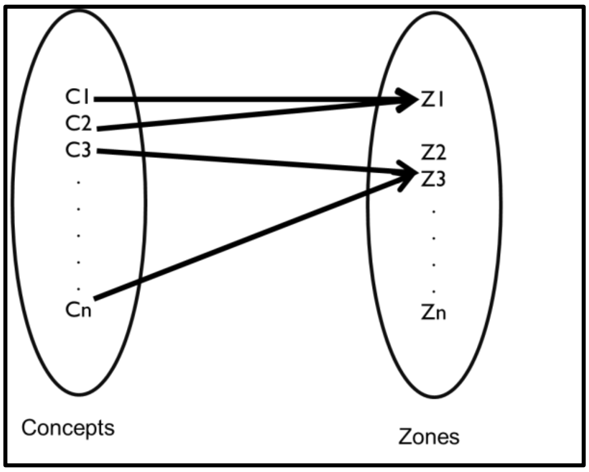

Figure 5: Venn Diagram of concepts mapped to zones

Documents and queries are represented as sets of terms. That is, each term is only considered present or absent in a document. Using the vector representation of the document above, the weightW_{ij}(∈ {0,1}) of term t_i, in document d_j is 1 if t_i appears in document d_j and 0 otherwise. Documents and queries are represented as sets of terms. That is, each term is only considered present or absent in a document. Using the vector representation of the document above, the weightW_{ij}(∈ {0,1}) of term t_i, in document d_j is 1 if t_i appears in document d_j and 0 otherwise.

$$f(x) = \begin{cases} 1, & t_i \text{ appears in zone gn} \\ 0\,, & \text{otherwise} \end{cases}$$

Proposed indexing technique is compared with linear search, inverted indices, zone based indexing and zone based indexing with n-ary tree. Zones are clusters with similar concepts. Tag based ranking is used here in zone based indexing. Tag based ranking is explained below.

Step 1: Firstly calculate tag based rank tbr

$tbr = \Sigma Weighting\ (w_i, t_i)$

For each occurrence of $w_i$the keyword is assigned a weighting factor

Step 2:In this step Drop XML tags.

Step 3: Write words and their positioning to the zone indexing in the database.

Step 4: Finally weighted zone indexing is calculated as:

$$\sum_{i=1}^{l} (g_i, t_i)$$

The tag based rank function for zone assigns value 1 if query exists in zone and zero otherwise. Weights are assigned to g_1, g_2, g_3...,g_n where n is the number of zones. Suppose g_1, g_2, g_3, g_4 are the four zones and we assign values to it 0.2, 0.3, 0.1, 0.4 respectively to the four zones so that weights add up to 1. The first, second and fourth zone add up more to the query. We would get our result from first, second and fourth zone reducing the time complexity of displaying the results from second and third zone which are contributing very less for our query.

5. COMPARATIVE ANALYSIS

Linear search takes more time to search in database so its RSV Value is low. Following diagram illustrates the RSV values of linear search; zone based indexing inverted indices and proposed system.

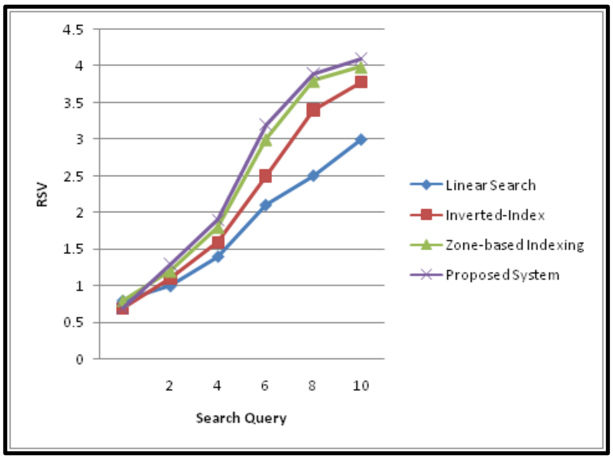

Figure 6: RSV Plots

Following diagram illustrates the time taken by Linear search, zone based indexing inverted indices and proposed system to give results to user.

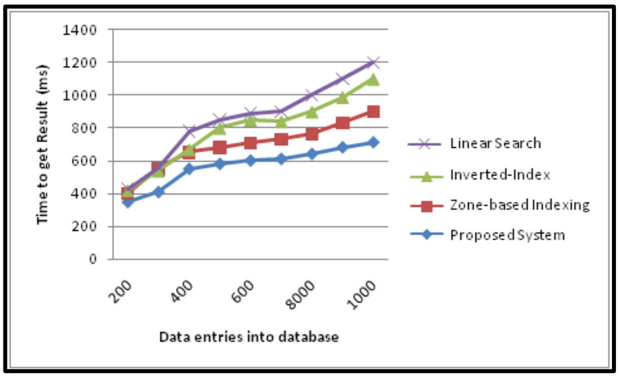

Figure 7: Retrieval time in microseconds

Aim of proposed system is give back results in less amount of time i.e low time complexity.

6. CONCLUSION

Aim of proposed system is to achieve high RSV value also at the same time giving the results that are expected to the user. If exact match is not found then proximity of keyword can be given to user i.e results that are nearer to or related to users query can be given. This proposed system is demonstrated on xml db. Proposed system can also be demonstrated on library. Even if exact match is not available then also some of the related books can be retrieved. In future the RSV value can be improved by optimizing the ontology database.

7. 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] 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. 287–291. Springer, Heidelberg (2010)

[3] Troels Andreasen, Henrik Bulskov,"Conceptual querying through ontologies", Fuzzy Sets and Systems 160 (2009) 2159 – 2172, Elsevier

[4] S.Geethalakshmi, S.Umamaheswari , "An Efficient Technique for Multikeyword based Search and Retrieval of Cloud Data", 2014 International Conference on Recent Trends in Information Technology.

[5] Robin Sharma, Ankita Kandpa,Priyanka Bhakuni,Rashmi Chauhan,R.H. Goudar,Asit Tyagi,"Web Page Indexing through Page Ranking for Effective Semantic Search",Proceedings of 7thInternational Conference on Intelligent Systems and Control (ISCO 2013).

[6] Santosh K. Vishwakarma, Kamaljit I. Lakhtaria, Divya Bhatnagar, Akhilesh K. Sharma,"An efficient approach for inverted index pruning based on document relevance",2014 Fourth International Conference on Communication Systems and Network Technologies.

[7] Vandana Dhingra, Komal Kumar Bhatia,"SemIndex: Efficient Indexing Mechanism for Ontologies "2014 IEEE

[8] Lachtar Nadia, "Design and implementation of information retrieval system based ontology", 2014 IEEE

[9] Rupali Chandsarkar, Radha Shankarmani, Prachi Gharpure "Information Retrieval System: for Skill set Improvement in Software Projects ",2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)

[10] Komal S. Mule, Prof. Arti Waghmare, "Review On Ontology Based Techniques In Information Retrieval Systems ", Multidisciplinary Journal of Research in Engineering and Technology, Volume 1,Issue 3, Pg.273-278

[11] Thomas R. Gruber, "A Translation Approach to Portable Ontology Specifications ", Knowledge Systems Laboratory Technical Report KSL 92-71

[12] Suruchi Chawla , Dr Punam Bedi, "Improving Information Retrieval Precision by Finding Related Queries with similar Information Need using Information Scent ", First International Conference on Emerging Trends in Engineering and Technology, 2008.

[13] Qing Chen, "Towards Web-based Information Retrieval in Grid Environment",Social Science Foundation of Hubei Province, IEEE,2010.