Suresh Gyan Vihar University, Jaipur
International Journal of Converging Technologies and Management (IJCTM)
Volume 2, Issue 2, 2016
ISSN: 2455 - 7528

An Analysis of Index Based Information Retrieval Algorithms

Deepika Sharma
deepika.sharma7@gmail.com
JagannathUniversity,Jaipur,Rajasthan,India
SandeepkumarPoonia
sandpoonia@gmail.com
JagannathUniversity,Jaipur,Rajasthan,India

Abstract:.From thousands of years humans store information in different ways. As from stone age to today's computer era there is a requirement to store information as well as retrieve this information in future and use it. Thus along with the storage there is a new requirement arises to retrieve the information also, not only accessing the information is important it also becomes relevant to the user. Nowadays with the advent of computer large amount of information can be stored easily and thus finding relevant information from such large amount of data become necessity. With this requirement their need some system which provides relevant and accurate information to the user according to the query given by the user such a system is known as Information Retrieval system which rely on textual keywords to index and retrieve documents. In IR system, indexing is a technique by which search of information become more accurate, fast and relevant. Index can be generated by the keywords present in the documents stored. Sometimes frequency of the result generated may also be stored in order of the frequency of the keywords presented in documents. This paper deals with analysis and comparison of different types of indexing techniques using different types of concepts and algorithms based on various parameter to find out their advantages and limitations for searching the relevant information.

I INTRODUCTION

As the volume of documents and thus information increases in the repository day by day there is a challenge to provide proper and relevant information to the user. Figure 1 shows a working of a typical information retrieval system.The motive behind this paper is to analyze the currently important algorithms for searching the relevant information.

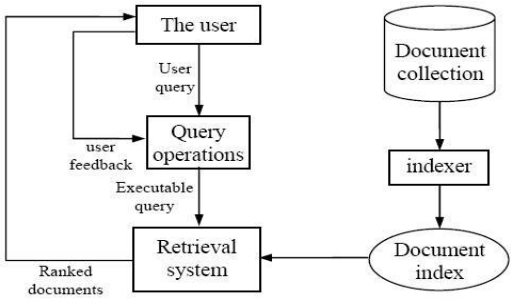

Figure1 Shows generation of documents according to user query with the help of index.

II RELATED WORK

ONTOLOGY BASED DOCUMENT INDEXING

Working of knowledge intensive organizations like consultancy services, consultancy and supply services, data processing services etc. are dependent

Suresh Gyan Vihar University, Jaipur
International Journal of Converging Technologies and Management (IJCTM)
Volume 2, Issue 2, 2016
ISSN: 2455 - 7528

on the information which has been gathered from various resources of information which may be inside the organization and/or external resources like the internet. This is very important to the organization to make the best use of gathered information from these sources. Knowledge intense organizations can do the knowledge management, which deals in the field of various activities relevant in knowledge life cycle: identification, acquisition, development, sharing, use and preservation of organization knowledge. Thus a system is designed for gathering the information for such tasks and the system developed is web based Webocrat system. Webocrat can interact at the knowledge level with the help of language. This language has been provided to the system by ontology with syntax and semantic rules.Use of ontology enables to define concepts and relations representing knowledge about a particular document in domain specific terms.

Scheme Of Document Retrieval

Here developed package with three different approaches to document retrieval: vector representation, latent semantic indexing method (LSI), and ontology-based method used in the Webocrat system.

Vector Representation Approach

This well knows approach is based on vector representation of document collection. First of all every document is passed through set of pre-processing tools (lower case, stop words filter, document frequency). Then a vector of index term weights is calculated as the document internal representation. These weights are calculated by most often used tf-idf scheme:

$$w = tf_{ij} idf_{ij}$$

where $tf_{ij} = \frac{freq_{ij}}{max_e freq_{ij}}$ and $idf_i = \log \frac{N}{n_i}$

$freq_{ij}$ is the number of occurrences of term t_i in document d_j, N is number of documents in collection, and n_i is the document frequency for term t_i in the whole document collection.

Such a vector is then normalized to unit length and stored into the term-document matrix A, which is internal representation of the whole document collection. In order to find some relevant document to a specific query Q it is necessary to represent the query Q in the same way as a document D_i (i.e. a vector of index term weights). Similarity between a query Q and a document D_i is computed as cosine of those two normalized vectors (document and query vectors).

$$Sim_{TF\text{-}IDF}(\vec{Q},\vec{D_i}) = \frac{\vec{D_i} \times \vec{Q}}{\|\vec{D_i}\| \, \|\vec{Q}\|}$$

Latent Semantic Indexing Approach

LSI approach is based on singular value decomposition of tf-idf matrix A. By this decomposition three matrixes are computed .

$$A = USV^T$$

where S is the diagonal matrix of *singular values* and U,V are matrices of left and right *singular vectors.* If the singular values in S are ordered by size, the first k largest values may be kept and theremaining smaller ones are set to zero. The product of the resulting matrices is a matrix approximatelyequal to A, and is closest to A in the least squares sense.

$$A \cong A_{SVD} \text{ where } A_{SVD} = U_k S_k V^T_k$$

Ontology Based Approach

Here for document retrieval ontology is used by Webocrat system. In Figure 2 the whole process of query processing by this approach is shown. When a user throw a query then first appropriate concepts are retrieved(here this work can be done manually) Then the set of concepts associated with each document is retrieved from database. Then using simple metric both the sets are compared which expresse[redacted]e similarity between a document D_i and given query Q.

$Q_{con} \cap D_{i,con}$ if $|Q_{con} \cap D_{i,con}| \neq 0$

$Sim_{onto}(\vec{Q},\vec{D}) =$

k

Where Q_{con} is a set of concepts assigned to query Q and D_{con}[redacted] a set of concepts assigned todocument D, and k is small constant, e.g. 0.1. Resulted number represents ontology-based similaritymeasure.

Suresh Gyan Vihar University, Ja
International Journal of Converging Technologies and Management (IJCTM)
Volume 2, Issue 2, 2016
ISSN: 2455 - 7528

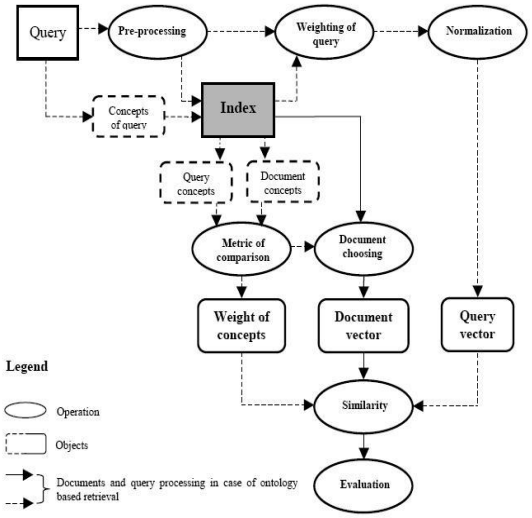

Figure 2 An ontology based document retrieval

KEYWORD BASED RETREIVAL SYSTEM

Here introduced a distributed information retrieval framework which is based on the Okapi probabilistic model, this is the framework which is able to achieve the same level of effectiveness as those achieved by a single centralized index system. Okapi model is proposed by Robertson and Spark Jones. This is the enhanced probabilistic retrieval model based on binary independence model. Here used a simplified Okapi weighting function, in which w_1 was assigned to a given term t_i in a document d and was computed according to the following formula:

$$w_i = (k_1 + 1) \cdot \frac{tf_i}{K + tf_i}$$

where

$$K = k \cdot (1 - b) + b \cdot \frac{l}{avdl}$$

l is the document length,
avdl is the average of document length.
b, k ,k_1 are constants.
tf_i is the occurrence frequency of term t in document d.
The following formula shows the weight given to the term t_i in document d.

$$qw_i = \frac{qtf_i}{k_3 + qtf_i} \cdot \log \left[\frac{n - df_i}{} \right]$$

INDEX BASED INFORMATION RETREIVAL SYSTEM

This is a three step process:

Indexing: In this step preprocessing of documents has been done. Here first a file is created for each document containing the words other than stop words (at, the, is, an etc.) also the stemming of the words is done, so that to get the words in their root forms (like: use is the root form of using, used, usable). In the next step frequency of each word has been counted and the words having frequency more than threshold value (based on a formula) is places as a index term. In the collected form all such terms create index table for that document.

Formulation: First step here is to expand the query based on domain knowledge stored in the form of ontological structure as a tree.

Second step is to apply the Preprocessing approach.

Comparison: The system compares the user query to the stored document representatives, and makes a classification decision about which documents to retrieve and in what order. Documents or parts of documents are displayed. Before searching user can select whether he wants to expand the query using tree or not. This comparison is carried out on the basis of matrix multiplication approach in which document representatives are converted into an id by term matrix and a matrix is generated for query terms. Multiplication of both provides necessary result to identify which document is more relevant to the query. Mathematically it can be shown as:

Suresh Gyan Vihar University, Jaipur
International Journal of Converging Technologies and Management (IJCTM)
Volume 2, Issue 2, 2016
ISSN: 2455 - 7528

Consider there are 2 documents (i and j) represented as:

Doc (i) = (Term (i1),Term(i2)……Term(ik)
Doc (j) = (Term (j1),Term(j2)……Term(jl)

Where k and l are no. of terms in respective documents.
So, all terms for all documents together can be represented as =
[Term (i1), Term (i2)……Term (ik) U (Term (j1), Term (j2)……Term (jl)]
- [Term (i1), Term (i2)……Term (ik) ∩ (Term (j1), Term (j2)……Term (jl)]

= [(Term (1), Term (2)……Term (n)]

i.e. Term(1)….Term(n)= all distinct terms of both documents i and j.

Here comparison is based on weighted values and implication of inverted document frequency (IDF):

Weight is how many times a term appeared in document. So weight implies how relevant the term is for that particular document

IDF is inverse of document frequency calculated for incorporating measure that favors terms which occur in fewer documents. The fewer documents a term occurs in, the higher this weight.

Thus this weight*IDF factor together will show a greater value if terms are important to document result. System has been tested on sample domain of computer science containing books of chapters and it is able to reduce the number of words to be searched in the file, thereby minimizing the search space. This effectively reduces the searching time as well. Reduction causes search spaces to be reduced more than 90% .

ATTENTION BASED INFORMATION RETREIVAL

It will be examined how attention data from the user can be exploited in order to enhance and personalize information retrieval. Up to now, nearly all implicit feedback sources that are used for information retrieval are based on mouse and keyboard input like click through, scrolling and annotation behavior. In this work, an unobtrusive eye tracker will be used as an attention evidence source being able to precisely detect read or skimmed document passages. This information will be stored in attention-annotated documents (e.g., containing read, skimmed, highlighted, commented passages). Based on such annotated documents, the user's current thematic context will be estimated. This context is increasingly taken into account in information retrieval systems .For instance; context can be generated implicitly or explicitly. One of the current challenges is to elicit the context of a user. This can be done explicitly, for example by asking the user, whether a document is currently relevant or not (i.e., explicit relevance feedback). To use such explicitly generated context is suggestive and yields better results in IR than without considering any user context. However, asking the user about explicit feedback requires a higher effort on the user's side and should therefore be avoided. Thus, implicit feedback recently gained in importance, i.e., observing the user's actions and environment and trying to infer what might be relevant for him. A very interesting new evidence source for implicit feedback is the user's eye movements, because mostly they reflect the user's visual attention directly. The eye trackers of today are unobtrusive and are able to identify the user's gaze with high precision. Therefore, applying an eye tracker as a new evidence source for the user's attention introduces a potentially very valuable new dimension of contextual information in information retrieval. It is clear that eye trackers will not be wide spread in the near future due to their expensiveness. However, if becoming less expensive, they might well be interesting for knowledge workers in middle- or large-sized enterprises. focus lies on local desktop and enterprise-wide search. As an eye tracker is not the only source for attention evidence, a model shall be developed, which integrates different attention evidence sources so that a standardized overall level of attention can be derived for any piece of text in a document.

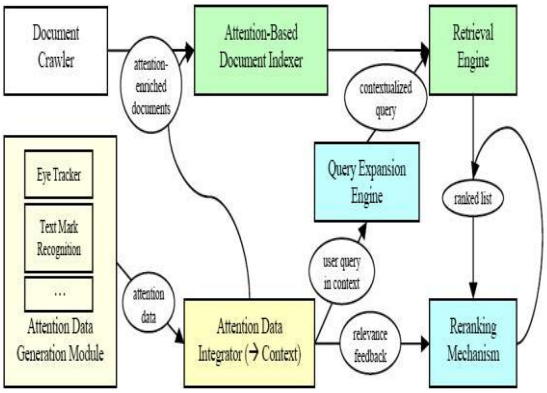

Figure 3 Shows generation of attention

Suresh Gyan Vihar University, Jaipur
International Journal of Converging Technologies and Management (IJCTM)
Volume 2, Issue 2, 2016
ISSN: 2455 - 7528

based index.

Attention-Based Index

The specific composition of a document reflects the mental models of the authors which might be very different from those of the readers. A reader of the document might only regard some parts of it with different intensities, namely those parts, which are of interest to him in his current thematic context (e.g., for his current task). Therefore, a document index for local desktop search, which supports retrieval of already used documents, should consider the user's degree of attention on the different parts of used documents. Secondly, there usually exists more than one document dealing with a specific topic being of interest to the user. Often, if a knowledge worker is making some document-based inquiry about some topic, he will not only regard one but several documents. To find these used documents at a later time (e.g., if a user remembers something interesting about a topic read some months ago), some connection between them should be maintained in the index.

RETRIEVAL EFFECTIVENESS OF AN ONTOLOGY-BASED MODEL FORCONCEPTUAL INDEXING

Traditional approaches in information retrieval employ keyword-based techniques to look for relevant data. This paper introduces a concept-based retrieval model, which tackles vocabulary mismatches through the use of domain-dependent ontologies. However, keyword-based searching is not always sufficient for retrieving the most relevant data, essentially because documents may convey desired semantic information even if they do not contain the exact keywords with the query. One way to alleviate the problem of retrieving relevant documents that are indexed with terms, which are superficially distinct but semantically equivalent to query terms, is to index documents according to their meaning rather than keywords (Woods, 1999). Indexing documents based on their semantics rather that their morphological content is known as *conceptualindexing*. This paper, propose a concept-based model for the index structure, which uses domain-dependent topical ontologies. With regard to converting words to meanings, the key element is to identify the concepts that characterize the thematic content of both documents and user issued queries. This paper describes a mechanism for the automatic identification of documents' thematic terms, which are employed by conceptual indexing module for representing the document's semantics at the index level. A critical feature in this technique is a disambiguation formula, which ensures that all document's thematic keywords represent concepts of a single thematic category. It also propose an automatic query semantics detection formula that helps the engines search mechanisms retrieve documents that highly correlate to the users' search intentions.Ontologies are define sets of representational terms, referred to as concepts, which are interrelated to describe a target world represented by links that are labeled so as to denote the type of relation that holds among concepts. There are two predominant approaches for building ontology; domain dependent and generic. Here domain dependent ontology is used for popular web topics In that respect we picked the top level categories of a popular Web directory, namely Dmoz1, and for each of these topics here developed a small ontology of concepts referring to the topic at hand.

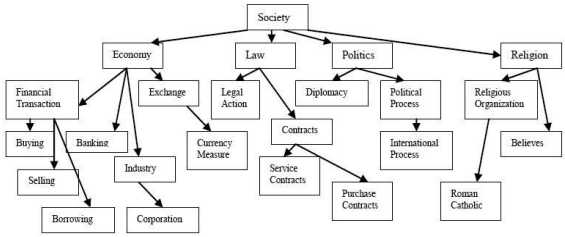

Figure 4 Shows a sample of our ontology for the Dmoz topic Society.

Finding Document Thematic Words now describe how explore the domain ontologies for converting the words found in a document into meanings. Thekey issue in converting words into meanings is to identify the appropriate concepts within a document that describe thedocument's thematic content. The computational model adopted here for finding the thematic words in a document relies onthe lexical chaining technique (Barzilay and Elhadad, 1997), which represents the content of a document as a sequence ofsemantically related terms When a document's term matches more than one concept in the ontology than apply a disambiguation formula so as to ensurethat the term will communicate to the generated chain its correct meaning. Concepts that are strongly associated to each other will be given high similarity scores based on their minimal distance from each other in the ontology's graph and their own correlation scores based on the number of common concepts that sub assume them in the ontology. Formally, for determining the

appropriate sense of a term matching multiple ontology nodes, rely on the matching nodes' correlation in the ontology graph, where the correlation score of a word pair w_1 and w_2 is formally determined as the product of the words' *Depth* (Song et al., 2004) and their conceptual similarity (*Sim*) (Resnik, 1995) in the enriched Greek WordNet. *Depth* of a word pair (w_1, w_2) is defined as:

$$\text{DepthScore}(w_1,w_2) = \text{Depth}(w_1)^2 \cdot \text{Depth}(w_2)^2$$

and words' similarity Sim (w_1, w_2) is determined by the set of Word Net concepts that subsume both w_1 and w_2 in any sense
of either words with a probability Pr:

$$\text{Sim}(w_1,w_2) = \max c \in \text{subsumers}(w_1,w_2) [- \log \Pr(c)]$$

Finally, we combine the above metrics to compute the Correlation Score between *w*1 and *w*2 that is formally defined as:

$$\text{Correlation}(w_1,w_2) = \text{Depth}(w_1,w_2) \cdot \text{Sim}(w_1,w_2)$$

Using the above formulas, document terms are disambiguated and participate in the document's lexical chain with the sense that has the highest correlation score to the senses of the other terms in that chain. Every chain element is associated with a particular meaning borrowed from the ontology. This way reduction in the document's content into a sequence of semantically related terms .Ontology-Based Document indexing intuitively, conceptual indexing is enabled through the organization of the collected documents into structures of thematic clusters. In this ontology-based model, a thematic cluster corresponds to the root node concept (i.e. the topic) of domain ontology and the structure of the cluster is represented by the ontology's hierarchies whose elements are specialized concepts of the root node concept. Based on the above, one way to address the conceptual indexing challenge is to view it from a hierarchical classification perspective. In other words, if here assign to every collected page an appropriate ontology topic than one can use the pages' lexical chain elements that match the nodes of the topic's hierarchies as indexing keywords for representing the pages' semantic content. To pick an appropriate category from the ontology's topics for representing a page's semantic content one can essentially need to identify the topical category (-ies) of the page's chain elements. To decide upon the topical category of a page's thematic words, here use the ontology's nodes that match the elements in the page's chain and than compute a score based on the number of words from the page's chain that associate to the ontology's topics (i.e. top level concepts). The calculation of that score, which represents the correlation between the page's thematic words and the ontology's topics, is formally defined as follows. Topic-Correlation score (Tscore): The T-score of a lexical LC_i for a particular ontology topic T_i is the number of elements of LC_i matched with subordinate concepts in T_i divided by thr total number of of elements in LC_i

$$\text{Tscore} = \frac{\#\text{ of elements of } LC_i \text{ matched}}{\#\text{of elements in } LC_i}$$

Based on the above formula, pick the topic that has the highest correlation with the elements in a page's lexical chain for representing the semantic content of that page. Following the process described above, one can assign to every page an appropriate ontology topic and then use the page's lexical elements that match the given topic hierarchies as the indexing keywords for the page.

III CONCLUSION

Based on the algorithm used, the indexing algorithm provides definite index to information retrieval system. A typical IR system should use indexing techniques based on the specific needs of the users. After going through exhaustive analysis of indexing algorithms. It is concluded that existing techniques have limitations particularly in terms of time response, accuracy of results, importance of the results and relevancy of results. An efficient indexing algorithm should meet out these challenges efficiently with compatibility with global standards.

IV REFERENCES

[1]. "Index based Information Retrieval System" Ambesh Negi1, MayurBhirud, Dr. Suresh Jain, Mr. Amit Mittal PG Scholar, IET DAVV, Indore Director, KCBTA, Indore Assistant professor, IET, IndoreInternational Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.2, Issue.3, May-June 2012 pp-945-948 ISSN: 2249-6645

[2]. "Term Proximity Scoring for Keyword-Based Retrieval Systems" Yves Rasolofo and Jacques SavoyUniversity de Neuchâtel, Neuchatel,Switzerland Published in Lecture Notes in Computer Science 2633, 1611-3349, 2003

[3]. "Ontology-based Information Retrieval" Jan ParalicDepartment of Cybernetics and AI, Technical University of Kosice,Letna 9, 04011 Kosice, Slovakia Ivan KostialDepartment of Cybernetics and AI, Technical University of Kosice,Letna 9, 04011Kosice, Slovakia

[4]."Attention-Based Information Retrieval" Georg BuscherGerman Research Center for Artificial Intelligence (DFKI)

[5]. "Retrieval Effectiveness OPfAn Ontology Based Model For Conceptual Indexing" Sofia Stamou Computer Engineering and Informatics Department, Patras University, 26500 GREECE