COMPARATIVE ANALYSIS OF VARIOUS INFORMATION RETRIEVAL TECHNIQUES

Poonam Jatwani, Govt. College for Women, Faridabad, Haryana, India(poonam.almadi@gmail.com)
Pradeep Tomar, Gautam Buddha University, Greater Noida, UP, India(parry.tomar@gmail.com)
Vandana Dhingra, Savitribai Phule Pune University, Maharashtra, India(vandana_dua_2000@gmail.com)

ABSTRACT

In today's world of globalization, World Wide Web has become an attractive commodity. It plays a vital role in everyone life. WWW can be used to perform various tasks like shopping, communicating with near and dear ones, for trading, entertainment and searching anything on internet etc. Information on the web is growing day by day, but this increase in volume of information has made the searching more difficult. To search specific information from sea of web documents has become very challenging as it shows many unwanted non relevant documents along with relevant documents. Intensive research has been going on developing methods that can retrieve relevant information. In this paper various information retrieval techniques are discussed and comparative analysis is presented. The study suggests that there is serious need to incorporate semantics and metadata to improve effectiveness of information retrieval. To retrieve relevant information semantic knowledge can be stored in the domain specific ontology which helps in understanding user's need to retrieve relevant information from the corpus.

INTRODUCTION

THE Semantic Web helps in extracting relevant information using many Information Retrieval (IR) techniques. Information Retrieval (IR) is defined as process of identifying and retrieving unstructured documents containing the specific information stored in them. Information Retrieval (Baeza-yates, R. and Ribeiro- Neto, B. ,1999) is finding material usually documents of an unstructured nature usually text that satisfy an information need from within large collections usually stored on computers. The objective of IR technique is to minimize the overhead of a user locating needed information. The overall goal of an information retrieval process is to retrieve the information relevant to a given request.

Current information retrieval tools mostly use keyword search, which mostly results in low precision and recall. Keyword search is not suitable to find relevant documents for a specific concept. In a traditional search engine the query terms are matched with the terms in an inverted index consisting of all the document terms of a text corpus. Only matched documents are retrieved and presented to the user. Users normally suffer from difficulties in finding accurate information on the web. The various issues related to IR are as follow:

- Traditional IR techniques are unable to handle big volume of text documents.
- Most of searching results display relevant as well as non relevant documents.
- Due to non relevant pages wrong ranking of web pages is presented.
- To extract relevant documents user have to spend time in searching both relevant as well as non relevant documents.
- Sometimes it becomes difficult to obtain useful information from the web documents.

EXISTING INFORMATION RETRIEVAL MODELS

Various information retrieval models exist to retrieve relevant information. Popular models are discussed below:

VECTOR SPACE MODEL: TF-IDF SCHEME

In vector space model both the query and each document are represented as vectors in the term space. A similarity coefficient that measure similarity among documents, and between documents and queries is used by a retrieval system to identify which documents are to be displayed to the user. Vector space model is based on exact word matching method, but exact word matching VSM Model does not support semantically word matching between queries and documents and thus suffering from the problem of polysemy and synonymy (P. Castells, M. Fernadez, D. Vallet, 2007). When different words were used in the input query this model failed to retrieve the relevant documents, which result in poor precision and recall.

LATENT SEMANTIC INDEXING APPROACH

The basic idea of LSI in information retrieval was proposed in 1988 by Scott Deerwester. Matrix computation is used as a basis for information retrieval in latent semantic indexing. A query-document similarity coefficient treats the query as a document and computes the SVD. This singular value decomposition (SVD) is used to filters out the noise found in a document, such that two documents that have the same semantics (whether or not they have matching terms) will be located close to one another in a multi-dimensional space (Rosario. B, 2000). However, the SVD is computationally expensive but it focuses on the need for a semantic representation of documents that is resilient to the fact that many terms in a query can describe a relevant document, but not actually be present in the document. Latent semantic indexing is the only strategy that directly addresses the problem that relevant documents to a query, at times, contain numerous terms that are identical in meaning to the query but do not share the same syntax. By estimating the "latent semantic" characteristics of a term matrix, LSI is able to accurately score a document as relevant to the query even though the term-mismatch problem is present.

CONCEPT INDEXING APPROACH

Concept based information retrieval is search for information objects based on their meaning rather than on the presence of keywords in the object . A content of an information object is described by a set of concepts in this model. Concept can be extracted from the text by categorization . Conceptual structure maps the descriptions of information objects to concept used in query. Conceptual structures can be general or domain specific, they can be created manually or automatically, they can differ in the form of representation and ways of constructing relationships between the concepts. Concept indexing uses centroids of clusters so called concept decomposition for lowering the rank of term-document matrix . In CI documents are presented as a linear approximation of concept vectors, terms are substituted by concepts, which are representatives of sets of terms (L. Shen, Y. K. Lim, H. T. Loh, 2004). In concept based informational retrieval systems, ontology can serve as a resource description and can be used for query formulation.

LANGUAGE MODEL

The basic idea of LM is given by researchers Ponte and Craft in 1998. The LM approach assumes that documents and expressions of information needs are objects of the same type, and assesses their match by importing the tools and methods of language modeling from speech and natural language processing. Language model work is based on word relationship inference. Zhai and Lafferty (Zhai, C. X. and Lafferty, J., 2004) proposed methods to retrieve information in language model using four types of queries: short keyword, short verbose, long keyword and long verbose.

HYPERSPACE ANALOG TO LANGUAGE MODEL

HAL is a cognitive - oriented model for representing word meanings in a high dimensional context space. Each word is denoted as a weight vector of its context words. The weights are calculated from the distance between a word and its neighboring words. A word closer to the target word is given a higher weight. Information flow provides a set of related words by combining the query words in the HAL space (Azzopardi, L., Girolami, M., and Crowe, M., 2005).Thus sense of a word and information flow inference for word relationships can be inferred from the neighboring context according to contextual information.

CONVENTIONAL HAL

In conventional HAL space (Amit Sheth & matthew Perry, 2008) the weights of a word vector are the co-occurrence values between the word and its neighbouring words. Neighbouring words are assumed to be independent of each other, ignoring temporal associations between neighboring words. Conventional HAL address the temporal association problem by relaxing the constraint on the independence assumption between neighbouring words.

EXTENDED PROBABILISTIC HYPERSPACE ANALOG TO LANGUAGE (EPHAL) MODEL

The epHAL incorporates the close temporal associations containing local and parallel remote dependencies between words in query documents to represent word occurrence relationships in a high-dimensional context space (Jui-Feng Yeh & Yu-sheng Lai, 2008). The information flow mechanism combines the query words in the epHAL space to infer related words for effective information retrieval.

CLUSTER- BASED MODELS

Cluster based retrieval group a set of documents into subsets or clusters in such a way that information in one cluster will be similar but they will completely different from the information in another cluster. It is assumed that each cluster is a representative of a particular topic. X. Liu and W. Bruce Croft in (Liu, X and Croft, W.B, 2004) proposed two cluster based retrieval, one for retrieving and other for using clusters to smooth documents. They used several TREC collections for evaluating these models and conclude that cluster based retrieval is feasible in LM framework.

COMPARISON OF INFORMATION RETRIEVAL TECHNIQUES

The Vector space retrieval strategy, assumes that terms are independent and ignore term associations. An inverted index is used to quickly compute the similarity coefficient. Each document in the collection does not need to be examined (unless a term in the query appears in every document) during the search. But with LSI, an inverted index is not possible as the query is represented as just another document and must, therefore, be compared with all other documents. LSI implementation needs additional storage space with more computing time. The premise is that more conventional retrieval strategies i.e. vector space, probabilistic and extended Boolean all show poor results due to exact match of keywords. Since the same concept can be described using many different keywords.

In LSI same concept can be described using different terms, so LSI solves the synonymy problem (J. Dobša, B. Dalbelo-Bašić, 2004) The LSI approach did not use stemming or stop words, so when the same terms were used for both methods, LSI was comparable to VSM and offers a partial solution for polysemy problem. SVD itself is computationally expensive and take up more storage space. LSI is a improvement of vector space model and opens a promising field of research by using term associations.

Concept indexing addresses the problem of synonymy and polysemy. Concept indexing is computationally more efficient and requires less memory then LSI. In case of LSI, documents are projected in the means of the least squares on the space spread by the first k left singular vectors while in CI, documents are projected on the space of k concept vectors. Concept vectors are sparse and can be labelled by terms, which have greatest weight in them. The reason for better interpretability of the CI method compared to LSI is in fact that concept vectors are more interpretable than singular vectors (Karypis G. , Hong & Han, 2000).

A proper representation of word meanings and an inference mechanism are needed for word reasoning. Word relationships can be explored using ontology based approaches, Language models and Hyper space analog to language based models. In ontology based model, semantically related words are derived from synonyms and hypernym-hyponym relationships. But in language model no relationship between terms are considered and no inference is involved. In language model, N-gram model is effective to explore local dependencies but ineffective for remote dependencies due to high computational complexity. The resulting model is mathematically precise, conceptually simple, computationally tractable, and intuitively appealing. It also seems necessary to move beyond a unigram model to accommodate notions of phrase or passage matching or boolean retrieval operators. Subsequent work in the LM approach has looked at addressing some of these concerns, including putting relevance back into the model and allowing a language mismatch between the query language and the document language (Bain, J., Song, P.D., 2005).

Song and Bruza [2001; 2003] adopted the Hyperspace analog to language (HAL) model to enhance the performance of short query information retrieval when keyword is absent using information flow for word relationship reasoning. HAL based models, such as conventional HAL and probabilistic HAL consider remote dependencies by the word independence assumption, but ignore temporal associations between words. The epHAL model develops a more generalized form in the HAL space than the conventional space.

Flat clustering is efficient and conceptually simple but its main drawback is it return a flat unstructured set of clusters, which require a pre specified number hierarchical of clusters as input and are nondeterministic. We select flat clustering when efficiency is important and we use hierarchical clustering to minimize the limitations of flat clustering like not enough structure, predetermined number of clusters, non determinism. Clustering can speed up search. Hierarchical clustering produces better clusters than flat clustering. Using cluster model, we find the

clusters that are closest to the query and select documents from these clusters. Within this smaller set, we can compute similarities exhaustively and rank documents in the usual way. Because there are many fewer clusters than documents, finding the closest cluster is fast, and because the documents matching a query are all similar to each other, they tend to be in the same clusters.

Fig. 2 summarizes the trends in which the search engines are enhancing/upgrading their capabilities by covering more and more semantic information and adopting better representation scheme.

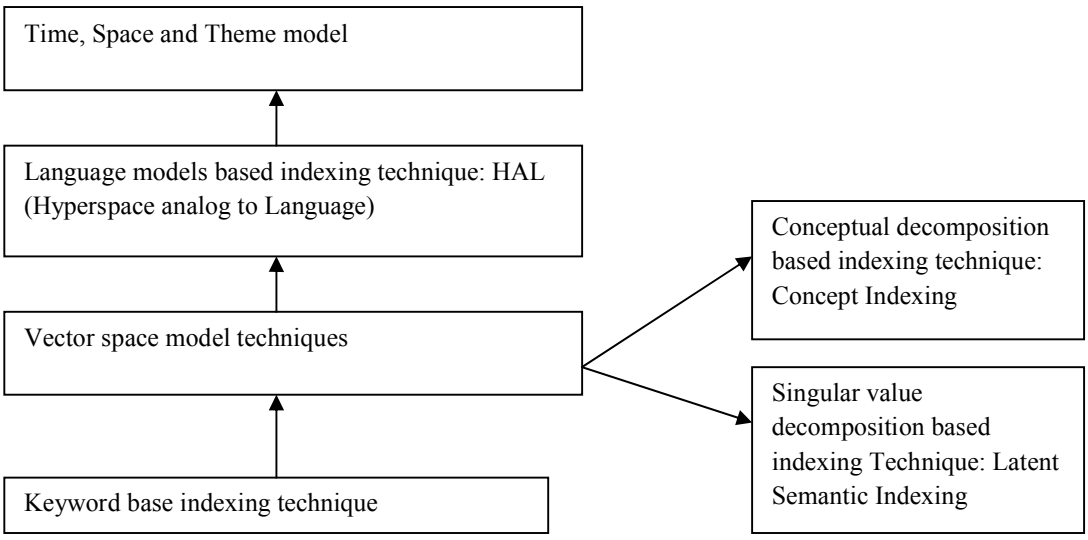

Figure2 : Trends in Semantic Indexing Techniques

ANALYSIS OF INDEXING TECHNIQUES

Table 1: Comparison of VSM, LSI, CI Models in respect of Parameters Precision, recall, Synonymy, Storage, Efficiency

Model/ Parameter	**Vector Space Model**	**Latent Semantic Indexing**	**Concept Indexing**
Precision	Poor	Improved Precision than VSM	Better Precision than LSI
Recall	Poor	Improved Recall than VSM	Better Recall than LSI
Synonymy	Synonymy Problem exist	Nicely deals with synonymy	Perfectly deal with Synonymy
Polysemy	Polysemy problem Exist	LSI offer a partial solution to Polysemy Problem	Perfectly deal with Polysemy Problem
Storage	Less Space is required	Requires additional storage space	Requires Less Space
Efficiency	Less efficient	Requires additional computing time than VSM.	It is more efficient than LSI

An experiment was conducted by (Jan Paralic, Ivan Kostial, 2003) for retrieval mechanism over VSM, LSI and ontology based retrieval. Experiment results showed that ontology based approach are very promising and provide better retrieval efficiency than VSM or LSI.

In general, ontology (Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. 1999) represents a conceptualization of a domain in terms of concepts, attributes and relations. In a conceptual model each concept is associated with a set of attributes. Ontology also defines a set of relations among its concepts. But extraction of the semantic concepts from the documents is the key issue using an ontology based model. There is need to identify appropriate concepts that describe and identify the documents.

Manually processing is difficult due to their increasing volumes. To generate semantic contents by mapping the information contained in existing web pages into concepts there is need to design Ontology based information retrieval Model which provides such a mechanism that no irrelevant concept will be retrieved and that relevant concepts will not be discarded.

CONCLUSION

In this paper, various information retrieval techniques are briefly discussed and then a comparative analysis of these is presented. The study suggests that current information retrieval techniques cannot exploit semantic knowledge within documents and give precise results. There is need to incorporate additional semantic information such as theme, time, space, relations among objects, ontological information among classes and their objects along with conceptualization in a cohesive manner. There is need to build a new paradigm for information retrieval that is compatible with all standards and provides effective and fast search. This work will focus on improving effective information obtaining mechanism using more effective inference mechanism for finding similar concepts to given query so that better precision and recall could be achieved.

REFERENCES

Amit Sheth and matthew Perry(2008). Traveling the semantic web through Space, Time, and Theme. *IEEE Internet computing, 12(2):. 81-86* .

Azzopardi, L., Girolami, M., and Crowe, M. (2005). Probabilistic hyperspace analogue to Language. *In Proceedings of the 2^{8th} Annual International ACM SIGIR Conference on Research and development in Information retrieval. ACM Press, New York, NY, 575-576.*

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic web. *Scientific American, 34-43.*

Baeza-yates, R. and Ribeiro- Neto, B. (1999). Modern Information Retrieval, Addison Wesley.

B.Rosario(2000). Latent semantic indexing: An overview. *Techn. Rep. Infosyss 240 Spring Paper, University of California, berkely.*

Bain, J., Song, P.D. (2005). Query expansion using term relationships in language models for information retrieval. *In proceedings of the 14th ACM international conference on information and knowledge management. ACM press New York ,688-695* .

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them ? *IEEE Intelligent Systms, 14(1), 20-26.*

George Karypis, Eui-Hong (Sam) Han(2000). Concept Indexing: Fast Supervised Dimensionality Reduction Algorithm with application to Document Categorization & Retrieval. *Proceeding of CIKM2000, ACM Press, pp-12-19* .

Jan Paralic, Ivan Kostial (2003) .Ontology –based Information Retrieval, *Proceedings of the 14th International Conference on Information and Intelligent System(ITS), Varazdin, croatia* .

J. Dobša, B. Dalbelo-Bašić(2004). Comparison of information retrieval techniques: latent semantic indexing and concept indexing, *Journal of Inf. And Organizational Sciences, 28(1-2): pp.1-17.*

Jui-Feng Yeh. Yu-sheng Lai. (2008).Extended Probabilistic HAL with close Temporal Association for Psychiatric Query Document Retrieval. *ACM Transactions on Information Systems, 27(1),Article 4* .

L. Shen, Y. K. Lim, H. T. Loh (2004) .Domain- specific Concept-based Information Retrieval System . *Proceeding 2004 IEEE International, 2(21-21) : 525-529* .

P. Castells, M. Fernadez, D. Vallet(2007). An adaption of the vector-space model for ontology-based information retrieval, *IEEE transactions on knowledge and Data Engineering, 19(2):261-272.*

Zhai, C. X. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inform. Syst. 22(2): 179–214.*