A Glance on Information Retrieval Techniques

Sudhir Soman
Assistant Professor, Dept. of computer science
Tilak Maharashtra Vidyapeeth, Pune.

Abstract- *Information retrieval is a very significant area of today's Information Technology world. There are two dimensions of information techniques, first dimension is mathematical basis [5] and second dimension is properties of models [5] available today. In mathematical basis, Set-theoretic models, Algebraic models, Probabilistic models, Feature-based retrieval models are involved. [5] And in properties of the model, models without term-interdependencies, models with immanent term interdependencies and models with transcendent term interdependencies are involved. [5]*

Keywords- Information retrieval, mathematical basis [5], properties of the model [5], first dimension, second dimension.

I. INTRODUCTION

Information is the most important key factor of every area of life. Every time users need to retrieve information for many reasons. This retrieved information should be relevant, precise and useful. Many researchers have done prominent work in the area of information retrieval with many aspects and perceptions. Information retrieval technique refers to obtaining relevant information from databases that consists of variable data stored at one or different places. [2] The information retrieval technique may not inform the user on topic of his enquiry request but it may inform the existence or non existence and whereabouts of databases corresponding to user's enquiry request. [6] But the retrieval technique must make sure that what is expected from the user should be received to him with utmost accuracy and precision.

1. The first dimension-Mathematical basis retrieval technique [5]

A. Set theoretic models [5]

The first dimension i.e. mathematical basis retrieval technique consists of set theoretic model which further consists of Standard Boolean model, Extended Boolean model and Fuzzy retrieval. [5]

Standard Boolean model is based on classical set theory and Boolean logic. [5]

In Extended Boolean model the term weights in queries is considered. The purpose is to overcome drawbacks of standard Boolean model. [5]

Fuzzy retrieval is based on fuzzy set theory and the extended Boolean model. [5] There are two fuzzy retrieval models Mixed Min and Max (MMM) [5] and the Paice model. [5]

B. Algebraic models [5]

Algebraic models consist of Vector space model, generalized vector space model, (enhanced) topic-based vector space model, extended boolean model and latent semantic indexing or latent semantic analysis. [5]

Vector space model is for representing text documents (and any objects, in general) as vectors of identifiers. [5]

Generalized vector space model is an extension and generalization of vector space model. [5]

Topic-based vector space model extends the vector space model by removing the constraint that the term vectors be orthogonal. [5]

Extended Boolean model was designed by Gerard Salton 1983, to overcome the drawbacks of the Boolean model. [5]

Latent semantic indexing or latent semantic analysis is a type of indexing and information retrieval method which uses a mathematical technique named singular value decomposition to find and identify patterns in between terms and concepts in text where that text is an unstructured collection. [5]

C. Probabilistic models [5]

Probabilistic models further consist of binary independence model, probabilistic relevance model , uncertain inference, language models, divergence from randomness model and latent dirichlet allocation. [5]

Binary independence model is one of the probabilistic information retrieval techniques which make some simple assumptions to prepare the estimation of document or query similarity probability feasible. [5]

Probabilistic relevance model is useful to derive ranking functions to rank matching documents according to their relevance; these functions are used by web search engines. [5]

Uncertain inference is way to formally define a query and document relationship in Information retrieval. It was first described by C. J. van Rijsbergen, 1986. [5]

Language model is actually termed as statistical language model which is a probability distribution over sequence of words. [5]

Divergence from randomness model is a type of probabilistic model which is one of the very first models in the field of information retrieval. This divergence from randomness model is used to test the amount of information carried in the documents. [5]

Latent dirichlet allocation (LDA) is a generative statistical model which allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. Latent dirichlet allocation is an example of a topic model . It was first presented as a graphical model for topic discovery by David Blei, Andrew Ng, and Michael I. Jordan in 2003. [5]

D. Feature based retrieval models [5]

Feature based retrieval models see documents as vectors of values of feature functions .They find the best way to combine these features into a single relevance score by learning to rank methods. [5]

2. The second dimension- Properties of the model. [5]

The second dimension i.e. Properties of the model consists of Models without term-interdependencies, Models with immanent term interdependencies and Models with transcendent term interdependencies. [5]

A. Models without term interdependencies [5]

Models without term interdependencies treat different terms or words as independent terms or words. [5]

B. Models with immanent term interdependencies [5]

Models with immanent term interdependencies allow a representation of interdependencies between terms. [5]

C. Models with transcendent term interdependencies. [5]

Models with transcendent term interdependencies normally allow a representation of interdependencies between terms. [5]

II. LITERATURE REVIEW

In this paper the authors have discussed about Cross-language retrieval systems, two sources of translation knowledge, implementation of six query translation techniques, better effectiveness by using machine translational systems, cross language information retrieval (CLIR), a strategy for document translation where automatic machine translation is used to translate each document in single language. [3]

In this paper the author has discussed about information retrieval indexing techniques. The paper tells about structures used in inverted files, use of sorted arrays, use of B-trees, use of Suffix Trees, signature files, a comparison of indexing techniques, i.e. performance Comparison, Stability Comparison, Limitations Comparison, a critical analysis of indexing techniques and benefits, limitations and challenges that are associated with the use of every technique.[2]

In this paper the authors have focused on Natural Language Processing (NLP).NLP techniques such as Hierarchical Conditional Random Fields (HCRF) and extended Semi-Markov conditional random fields (Semi-CRF) along with Visual Page Segmentation is used to get the accurate results. Database wrapper based methods have a problem of context switching. Very less work has been done for integrating web NLP techniques and entity extraction. The design system is implemented based on text content which is analyzed through user perception and the web page layout understanding. The author claims that 74.6% and 86% is achieved for the web page and the input text respectively. [9]

In this paper the authors are discussing about the results of a survey conducted by them to determine the information retrieval techniques used by the teachers and students of Midnapore College library. These papers examine the result from a questionnaire based survey conducted at the library. The study reveals that a significant number of users search information regarding the library material through OPAC (Online Public Access Catalog) despite encountering problems. According to the authors the readers use either

metadata or a search engine. Metadata provides less but relevant information whereas search engines provide lots of information but some of it is unusable. So authors suggest that the use of search engines should be appropriate so that they will get required information timely and economically. [1]

In this book they have focused on Boolean retrieval , vocabulary, postings lists ,dictionaries ,tolerant retrieval , index construction , Index compression , Scoring, term weighting , the vector space model , Computing scores in a complete search system ,evaluation in information retrieval ,relevance feedback and query expansion , XML retrieval , probabilistic information retrieval , language models for information retrieval , text classification , Naive Bayes , vector space classification , support vector machines ,machine learning on documents , flat clustering , hierarchical clustering , Matrix decompositions, latent semantic indexing , web search basics , web crawling ,indexes and Link analysis.[4]

In this book they have discussed about representation of text document inside a computer, automatic classification methods in information retrieval, file structures in the viewpoint of information retrieval, search strategies, probabilistic retrieval to enhance its effectiveness, attempted to provide a foundation for theory of information system evaluation, guessing about future information retrieval techniques and the areas of research where work is needed. [8]

REFERENCES

[1] Avijit Dutta and Subarna Kumar Das, Information Retrieval Techniques used by the Midnapore College (Autonomous)Library Users: A Study, 9th Convention PLANNER-2014
Dibrugarh University, Assam, September 25-27, 2014© INFLIBNET Centre, Gandhinagar

[2] Zohair Malki, Comprehensive Study and Comparison of Information Retrieval Indexing Techniques, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, 2016

[3] Douglas W. Oard, D. Farwell et al. (Eds.): AMTA'98, LNAI 1529, pp. 472 483, 1998.Springer- Verlag Berlin Heidelberg 1998.

[4] Christopher D. Manning, Stanford University, Prabhakar Raghavan, Yahoo! Research, Hinrich Schutze, University of Stuttgart, Introduction to Information Retrieval, ISBN 978-0-521-86571-5 hardback, Cambridge University Press 2008.

[5] Https: // en.wikipedia.org/ wiki/ Information retrieval

[6] Lancaster, F.W., Information retrieval systems: Characteristics, Testing and evaluation, Wiley, New York (1968).

[7] C. J. van RIJSBERGEN B.Sc., Ph.D., M.B.C.S. INFORMATION RETRIEVAL, http:// openlib.org/ home/ krichel/ courses/ lis618 / readings/ rijsbergen79_infor_retriev.pdf

[8] http:// www.dcs.gla.ac.uk/ Keith/ Chapter.1/ Ch.1.html#[2]

[9] Rini John and Sharvari Govilkar, Information retrieval technique for web using NLP, International Journal on Natural Language Computing (IJNLC) Vol. 6, No.5, October 017 DOI: 10.5121 / ijnlc. 2017. 65011

[10] Charles H. Heenan, A Review of Academic Research on Information Retrieval, Regnet Project, Department of Civil and Environmental Engineering Stanford University, Stanford , California 94305, the National Science Foundation, under Grant No. EIA-0085998, August 6, 2002.

[11] Mark Sanderson, W. Bruce Croft, The History of Information Retrieval Research, Proceedings of the IEEE • 2012