ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 3, May 2012

Optimization of Information Retrieval by Advance Indexing Models and FIRS

Divya Jyoti Shrivastav[1], Waseem Ahmad[2]

Abstract- **Today Internet is becoming very wide and user want to retrieve the exact information related with the query. Sometimes user query result return by the search engine is not exactly same. For this type of information retrieval problem may be reduced by underlying Indexing Techniques namely Latent Semantic Indexing, Latent Semantic analysis, PLSI, and Latent Dirichlet Allocation Model for provide semantic tools for text and also describe some multimedia retrieval tools to provide accurate precision and remove low recall rate .Here we introduce a new model for linking document called as LTHM and Fuzzy Information Retrieval System(FIRS) for describe the weight based queries. Basically PLSI and LDA are used for the contents of plaintext so for web text LTHM is used to describe the hyper linking of document with web. The prospective area of information retrieval where FIRS is needed is described with justification. FIRS are used to define the accuracy and precision of web documents and queries.**

Keywords: **Latent Semantic Indexing, PLSA, LTHM, FIRS**

I. Introduction

Information retrieval systems are designed to solve the problem of extraction of exact information from the huge collection of web documents. The main aim of Information Retrieval is relevant information with accuracy and precision. This is a very complex task for traditional Information system to find the uncertainty in queries, Imprecision and subjectivity. Problem with the existing information Retrieval system is text classification and matching documents. This dramatic increase in available information has driven the need for the development of *automatic* information retrieval. A Fundamental deficiency of traditional IR is that the keyword searchers use does not get the information they indexed. Some issues has been described as Polysemy (More term may have different meanings in a query) and Synonymy (The same meaning can be expressed by two or more different terms). At present, the soft computing tools include fuzzy sets, Conceptual Fuzzy logic, and artificial intelligence with neural network. Fuzzy set provide useful categories associated with properties as 'large','small','narrow','very narrow' possibly modulated by linguistic hedges. The grammatically of sentences,synonymy,hypermony,semiotics of text also involve fuzzy notions.Rieger[11-13] for the development of fuzzy semantic information retrieval also view the text based in co-occurrence of words. Semantic Indexing is a way to recover this issue by providing different model that is Latent Semantic Indexing, Latent Semantic analysis, PLSI, and Latent Dirichlet Allocation Model and a new model for combining the PLSA and LDA model.

The rest of this paper is organized as follows: Section II describe the characteristics of information retrieval models and the limitation of existing model. Section III provides an introduction to fuzzy set and semantic indexing Models. Section V covers, in detail, the use of FIRS and proposed new indexing models and tools for the development of Information retrieval system. Sections VI provide the conclusion and future scope of research in the area of Web information retrieval.

II. Information Retrievals Models and its limitation

Mapping functions are the main engine of information retrieval systems. Once representations for documents and queries are built, these representations are used by the matching function to achieve the three following related tasks:

1. To locate or identify items related to a user query.
2. To identify both related and distinct documents in the collection.
3. To predict the relevance of a document to the user's information request through the use of index terms with well defined scope and meaning.

Many matching functions have been proposed over the years by researchers in the information retrieval area. In this section we examine three different matching function models, namely the *Boolean*, v*ector space* and *probabilistic* models.

A. *Boolean Model*

The *Boolean mode l* is considered to be the simplest matching function in information retrieval. Relationships or similarities between individual documents are not utilized, neither are any relationships between query terms. In systems which use the Boolean model, the users' query is represented only as combinations of terms that a relevant document is expected to contain. For example, one may require all documents which contain the two terms (*manufacturing* and *implementation*) or the three terms (*cheap*, *material* and *good*). The query *Q* can be formulated as

Q = (manufacturing **AND** implementation) **OR** (cheap **AND** material; **AND** good)

Simplicity of implementation is the main advantage of Boolean model. The main drawback of Boolean model is that is it concentrates on building a retrieval model that has the ability to find the weighted relevant documents against the query of user.

B. *Vector Space Model*

The vector space model represents both the queries and web documents in the term of vector. Both document and query are described as T dimensional space or 3D space Where T represents the unique terms in the document collection. Fig.(1) shows an example of a vector space model representation for a system with two documents.

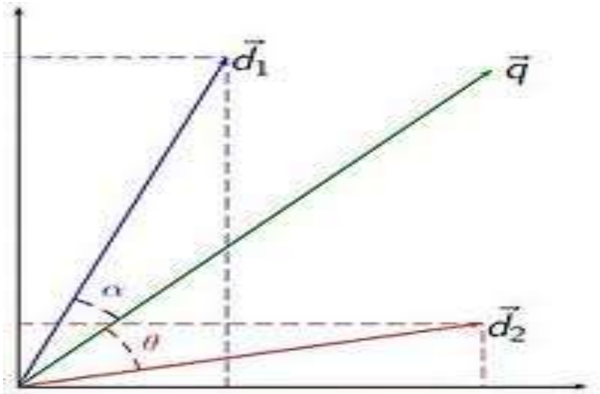

Fig.1. Vector Space Model

Each axis in the space describes different term. A similarity computation measuring the similarity between a particular document vector and a particular query vector as a function of the magnitudes of the matching terms in the respective vectors may be used to identify the relevant documents. The simplest such scheme to calculate the similarity is to assume that the document containing the most terms from the query will be the most relevant. Thus the similarity between a query *Q* and the *kth* document, *Dk*, can be calculated as an *inner product* of term vectors in *Q* and *Dk*. Formally it can be represented a similarity measure

$$\mathrm{sim}(Q,D_k) = \sum_{i=1}^{n} q_i\, t_{ik}$$

where

***Q* is the query vector**

***D_k* is the *k^{th}* document vector in the collection**

***q_i* is the term *i* in the query Q**

t_{ik}* is the term *i* in the document *Dk

***n* is the total number of query terms.**

Besides this approach, cosine correlation function also describes the vector similarity measure. In the cosine correlation function, the angle between documents or documents and a query measures the similarity between the vectors that represents them. The similarity between d1 and q would be measured by the angle $\tilde{\alpha}$The similarity between documents d2 to query Q is measured by angle θ. The cosine correlation function is Refer to “(1)”

$$\text{Cosine correlation} = \frac{\sum_{i=1}^{n} q_i\, t_{ik}}{\left(\sum_{i=1}^{n} q_i^2\right)^{1/2} \left(\sum_{i=1}^{n} t_{ik}^2\right)^{1/2}} \qquad (1)$$

The lack of formal methods to support the vector space model in handling uncertainty has driven research in information retrieval towards seeking models that can support uncertainty.

C. Probabilistic Model

The probabilistic model attempts to address the uncertainty problem in information retrieval through the formal methods of probability theory. Unlike in the vector space model, in this model the document ranking is based on the probability of the relevance of documents and the query submitted by the user. This has been formalized and is known as the *Probability Ranking Principle*. There are three different models of probabilistic retrievals: *binary independence* .the *unified model* [Roberston14] and *retrieval with probabilistic indexing* .The models differ in their treatment of and assumptions behind the *probability of relevance*. In this section, we analyze the formulation of these probabilistic models and state the assumptions associated with them.

We have stated above that the probabilistic model assumes that the terms in the document are distributed independently. However, Rijsbergen [Harper15] argues that this assumption is often made as a matter of mathematical convenience, although it is generally agreed that exploitation of associations between items of information retrieval systems, such as index terms or documents will improve the effectiveness of retrieval.

D.Page Ranking:

Page ranks are important since human beings find it difficult to scan through the entire list of documents returned by the search engine in response to his/her query. Rather, one sifts through only the first few pages, say less than 20, to get the desired documents. Therefore, it is desirable, for convenience, to get the pages ranked with respect to “relevance” to user queries. However, there is no definite formula which truly reflects such relevance in top-ranked documents. The scheme for determining page ranks should incorporate 1) weights given to various parameters of the hit like location, proximity, and frequency;

2) weight given to reputation of a source, i.e., a link from yahoo.com should carry a much higher weight than a link from any other not so popular site; and 3) ranks relative to the user.

III. Use of Fuzzy Logic and indexing Models in Information Retrieval System.

An IR System aims to evaluate users Queries The result produced by a query evaluation is the set of documents estimated relevant to the information. To provide relevant Information IR uses following semantic indexing models

A. Latent Semantic Indexing (LSI)

Latent Semantic Indexing (LSI), one kind of LSA model, was firstly proposed to address finding semantic relevance in the context of information retrieval and digital library. Researchers utilized it to identify the semantic themes hidden in a large amount of document collections.

LSI algorithm has achieved great success in text mining and has been extended to other related applications . The standard LSI algorithm is based on a SVD operation.

Latent Semantic Indexing (LSI) is an approach to capture the latent or hidden semantic relationships among co-occurrence activities. In practical applications, *Singular Value Decomposition* (SVD) or *Principal Component Analysis* (PCA) algorithms are employed to generate a reduced latent semantic space, which is the best approximation of the original input space and reserves the main latent information among the co-occurrence activities, LSI has been widely used in information indexing and retrieval applications, Web linkage analysis and Web page clustering

B. Probabilistic LSA (PLSA)

PLSA model is an extension of Traditional LSI model, is used for Image Classification, Text analysis, and analysis of two mode data in information retrieval co-occurrence data related with applications. With this model it defines a proper generative model of data. Let the occurrence of term *w* in a document *d* be an event in model and z denote a latent variable associated with each event in the model. the generative process of this model is described as follows:

1. Choose a document over a distribution P(d_i)
2. Choose a latent topic z with probability P(z_k | d_i)
3. Choose a term w according to P(w_j| z_k)

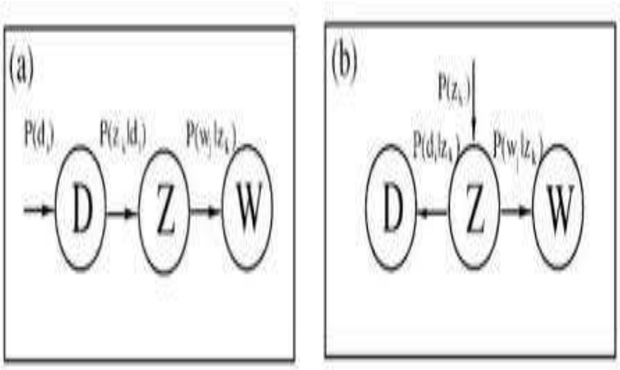

Fig. 2. Probabilistic Model”(a)” & “(b)”

The graphical model for this generative process is shown in the fig.2 (a). The PLSA model postulates that a document and a term w are conditionally independent given an unobserved topic z.The probability of an event P(w_j | d_i) is defined as (2)

$$P(w,d) = P(d)\ \Sigma_{z}\ P(w|\ z)P(z\ |\ d) \qquad (2)$$

The parameters of PLSA model are estimated by the latent variables models i.e. Expectation maximization (EM) algo, the computation of EM alternates between Expectation (E) step and the Maximization (M) step. During E step posterior probabilities of the latent variable z and M steps update the parameters as follows

E- step

$$P(z|w,d) == \frac{P(w|z)P(d|z)P(z)}{\sum_z P(w|z)P(d|z)P(z)}$$

M - step

$$P(z) = \frac{\sum_d \sum_w n(d,w)P(z|w,d)}{\sum_d \sum_w n(d,w)}$$

$$P(w|z) = \frac{\sum_d n(d,w)P(z|w,d)}{\sum_w \sum_d n(d,w)P(z|w,d)}$$

$$P(d|z) = \frac{\sum_w n(d,w)P(z|w,d)}{\sum_d \sum_w n(d,w)P(z|w,d)}$$

The graphical model for symmetric parameterization used in the model fitting process describe above is shown in fig. 2(b).The PLSA model does not explicitly specify how the mixture weights of the topic P(z) are generated ,making it difficult to assign probability to a new as yet unseen document. For Image retrieval we proposed a new method that is **C-PLSA** (Correlated PLSA model), that define the image correlation when estimating the parameters which often leads to inaccurate latent topics.

C.Latent Dirichlet Allocation

LDA model is a recently emerging generative model, which reveals the intrinsic correlation among co-occurrence via a generative procedure. It is a Bayesian based model which It remove the over fitting problem of PLSA. The discovered usage knowledge is then used to predict user potentially interested Web contents. The common strength of the latter two models is the capability of capturing the aspect space that associates with the discovered usage knowledge in addition to usage pattern mining itself.

In statistics, **latent Dirichlet allocation (LDA)** is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. The generative process of LDA model is follows:

1-For each topic z= 1----k choose W dimensional $\boldsymbol{\beta}_{\mathbf{z}}$ ~ Dirichlet (η)

2-For each document d=1……D Choose k Dimensional θ ~ Dirichlet (α)

-For each position i =1…..N_d

-Choose a topic $z_{i\ \sim}$ Mult(. | θ_d)

-Generate a term $w_{i\ \sim}$ β_{zi})

The procedure for inference under the LDA involves the computation of the posterior distribution of the hidden variables given a document:

$$\mathbf{P(\theta,z|w,\alpha,\beta)} = \frac{\mathrm{P}(\boldsymbol{\theta},\mathrm{z,w}|\boldsymbol{\alpha}\ ,\boldsymbol{\beta})}{\mathrm{P(w}|\ \boldsymbol{\alpha},\ \boldsymbol{\beta})} \qquad (3)$$

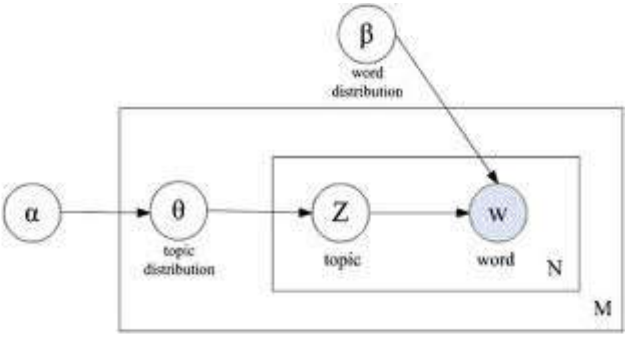

Fig.3. LDA Model

For example, an LDA model might have topics that can be classified as **CAT** and **DOG**. However, the classification is arbitrary because the topic that encompasses these words cannot be named. Furthermore, a topic has probabilities of generating various words, such as *milk*, *meow*, and *kitten*, which can be classified and interpreted by the viewer as "CAT". Naturally, *cat* itself will have high probability given this topic.

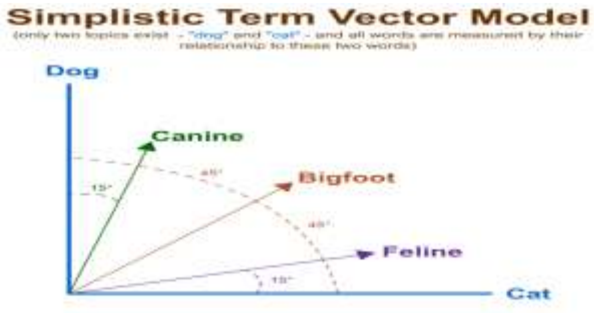

Fig.4. Simplistic Term Vector model

E.FUZZY SET

Information retrieval involves two finite crisp sets, a set of recognized index terms,

$$X = \{x_1,x_2,\ldots\ldots\ldots x_n\}$$

and a set of relevant documents,

$$Y = \{y_1, y_2, \ldots, y_n\}$$

In fuzzy information retrieval, the relevance of index terms to individual documents is expressed by a fuzzy relation,

$$R = X \times Y \text{------>} [0,1],$$

such that membership value R (xi, yj) specifies for each xi∈X and yj∈Y the grade of relevance of index term xi to document yj.

IV. COMMERCIALLY AVAILABLE SYSTEM FOR FUZZY IR

Here we list some commercially available tools

- Nzsearch: (www.searchnz.co.nz/) is a search engine based completely on FL. It considers the entire phrase rather than individual words for the purpose of matching. It also uses a "fuzziness" parameter while searching, which can be chosen from the set: "minimal," "normal," "moderate," "very," and "extremely."
- Finder: (www.finder.co.uk) uses "multidimensional optimization" to display the best or most suitable matches to a query unlike most existing search engines which display only the exact matches to a given query. Finder goes way beyond the simple "yes" or "no" criterion, used by most database query engines such as SQL or Btrieve. Finder uses scoring modules designed especially for each data type. If one is looking for a scarlet car, and the car in the database was cherry red, it would not ignore the entry altogether, it would just give it a lower score. The search engine looks at each element of the database and scores it using one of its knowledge-based scoring modules.

V. .PROPOSED SEMANTIC INDEXING MODEL

A. LATENT TOPIC HYPERTEXT MODEL (LTHM):

This model defines the extension of Link LDA-PLSA Model by defining the model by considering links between documents. It allows the existence of links from every web document to every web document, including self reference. The probability of generate a document k to a document k^n it depends on 1)The word topic from which is link is established(2)the in degree of the final target k^n (3) the mixture of topic of the target web document k^n.

Basically LTHM is a two step process. In the first process document content is created by LDA in which and in the second stage link is generated.

B.FUZZY INFORMATION RETRIEVAL SYSTEM:-

In the FIRS approach, work on IR involves designing sophisticated IRSs based on fuzzy logic and evolutionary computing with two different goals: facilitating the users the expression of their information needs in the form of advanced fuzzy queries, and effectively solving those needs with relevant information.

Design of FIRS incorporating an advanced query language able to manage multi-level, multi-granular and no-balanced linguistic information. Here we develop a new Retrieval system which defines the TF-IDF frequency and respond the weighted based queries after that the rules should be define on the basis of 'AND','OR' and 'NOT' operation.

.

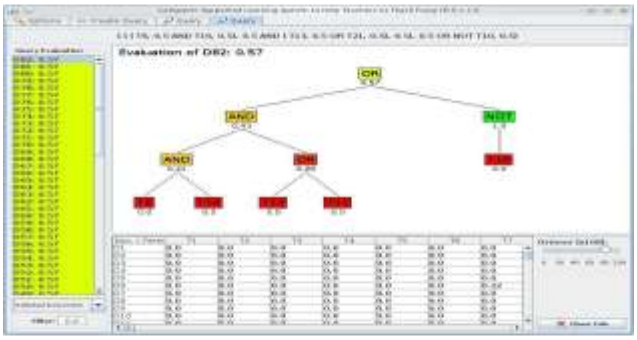

Fig.5. Fuzzy Information Retrieval System(FIRS)

VI. CONCLUSION

Information retrieval (IR) is used to find material an unstructured nature that satisfies information need from within large collections (usually stored on computers)

In information retrieval, documents and queries are represented by index terms. But the existing IR model does not provide the knowledge about the information, it has some uncertainty, lack of "approximate reasoning", Imprecision about the query. As a result, document and query. Thus, an Soft computing approach may be considered as a solution for the above problem in text retrieval task. It provides methods of representing documents and users' information

Three kinds of semantic analysis models, namely standard LSA, PLSA and LDA, are proposed to address to remove the problem of existing IR model. The proposed models provide the link analysis of web documents with the combination of PLSA and LDA and Fuzzy Logic remove the uncertainty problem of documents. Future scope is to reranked the indexing term by using different types of soft computing methods. Also the effect of different HTML tags will be defined quantitatively. And a probabilistic equation of user satisfaction will be derived.

.

ACKNOWLEDGEMENTS

For this research work I acknowledge the role of parent

Institution, Referenced research work and the valuable work of authors which if I missed to address and my friends.

REFERENCES

[1] Michael A. Casey, "Reduced-rank spectra and minimum entropy priors as consistent and reliable cues for generalized sound recognition," in *Proceedings of Euro speech*, 2001.

[2] Milind R. Naphade and Thomas S. Huang, "A probabilistic framework for semantic video indexing, filtering and retrieval,"*IEEE Transactions on Multimedia, special issue onMultimedia over IP*, vol. 3, no. 1, pp. 141–151, Mar. 2001.

[3] G. Iyengar and A. B. Lippman, "Models for automatic classification of video sequences," in *Storage and Retrieval from Image and Video Databases*. Jan 1998, vol. VI, SPIE.

[4] Schmidt, Thomas and Wörner, Kai (2009). "EXMARaLDA –Creating, analyzing and sharing spoken language corpora for pragmatic research." In: *Pragmatics 19*.

[5] Schmidt, Thomas and Bennöhr, Jasmine (2008). "Rescuing Legacy Data." In: *Language Documentation and Conservation 2*, 109-129.

[6] A. I. Joseph, I. Thomas-Kerr, S. Burnett, C. H. Ritz, S. Devillers, D. De Schrijver,and R. V. Walle. Is that a fish in your ear? A universal metalanguage for multimedia. IEEE Multimedia, 14(2):72{77, 2007.

[7] C.Ding, *A Similarity-based Probability Model for Latent Semantic Indexing,* Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, p.p-59–65, 2004.

[8] R.Price, and A.Zukas, Application of Latent Semantic Indexing to Processing of Noisy Text, Intelligence and Security Informatics, Lecture Notes in Computer Science, Vol. 34, Springer Publishing, p.p .602–603, 2007.

[9] Hou, J. and Y. Zhang, *Effectively Finding Relevant Web Pages from Linkage Information.* IEEE Trans. Knowl. Data Eng., 2003. 15(4): p. 940-951.

[10] B.Bartell,G. Cottrell,, and R.Belew, Latent Semantic Indexing is an Optimal Special Case of Multidimensional Scaling, Proceedings, ACM SIGIR Conference on Research and Development in Information Retrieval, p.p-161–167, 1992.

[11] B. B. Rieger, Feasible fuzzy semantics, *Proc.7th Inter. Conf. on Computa. Linguis.* (COLING78), Bergen, (Heggstad, K., ed.), 41-43, 1978.

[12] B. B. Rieger, The baseline understanding model.A fuzzy word meaning analysis andrepresentation system for machinecomprehension of natural language, *Preprints* 6th Europ. Conf. on Artif. Intellig. (ECAI/84),Amsterdam, (T. O'Shea, ed.), 748-749, 1984.

[13] B. B. Rieger, Computing granular word meanings. A fuzzy linguistic approach to computational semiotics, In Computing with Words (P. P. Wang, ed.), John Wiley & Sons, New York, 147-208, 2001.

[14] Robertson, S.E., Maron, S.E. and Cooper, W.S. Probability of Relevance: A Unification of Two Competing Models for Document Retrieval. *Information Technology: Research and Development,* 1(1):1-21, 1982.

[15] Harper, D. and van Rijsbergen, C.J. An Evaluation of Feedback in Document Retrieval Using Co-occurrence Data, *Journal of Documentation*, 34(3):189-216, 1978

[16] O. Etzioni and O. Zamir, "Web document clustering: A feasibility demonstration," in Proc. 21st Annu. Int. ACM SIGIR Conf., 1998,pp.46-54 .

[17] Klir, G.J. and Yuan, 60, "Fuzzy Sets and Fuzzy Logic Theory and Application"

[18] C. V. Negotia, "On the notion of relevance in information lnformation retrieval,"*Kybernetes*, vol. 2, no. 3, pp. 161–165, 1973.

[19] Ramesh Nallapati and William Cohen. 2008. Link- PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *International Conference for Webblogs and Social Media*.

[20] Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. 2008.Latent Topic Models for Hypertext. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*.

[21] Thomas Hofmann. 1999. Probabilistic Latent Semantic Analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*.

.

Divya Jyoti Shrivastav^1 received the B.Tech. Degree in computer science from B.I.T, Muzaffarnagar, India, in 2007.

Currently, she is a Senior Lecturer in N.I.E.T, Gr.Noida His research interests are in the area of data mining and knowledge discovery, pattern recognition, learning theory, and soft computing. Her three papers was published in National Conference.

Waseem Ahmad^2 received the B.Tech. & M.tech degree in computer science from AFSET, Faridabad India. Currently, she is a Senior Lecturer in AFSET, Faridabad. His research interests are in the area of Image Processing & Compression, data and Web mining, pattern recognition, learning theory, and soft computing.