International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

# A Survey on Application of Information Retrieval Models Using NLP

Dr.S.Gomathi[1], Dr. M. Lavanya[2]

[1,2]*Computer Science Department, SDNB Vaishnav College For Women Chromepet*

*Email:* [1]*gomathiganesh1978@gmail.com*, [2]*lavanyalalith1979@gmail.com*

***Abstract: Text mining accustomedestablish the hidden patterns and to get a vital data fromthe great deal of unstructuredknowledgeit's going to written material also. Different field techniques like machine learning, visualization, text analysis, database technology, statistics, knowledge management, natural language processing are incorporated in text mining. It's thought about to be mostly booming area next to big data and artificial intelligence. There exist varied techniques, tools and applications in text mining. Information retrieval thought-about being aeminent application of text mining. Information retrieval could be a key technology for knowledge management. Information retrieval is the method ofgetting and presenting additionalconnected data from the biggest assortment of knowledge resources in step with to the user's need.This paper is a survey of discussing concerning varied application of Information Retrieval using NLP and available methods of information retrieval. This survey discussesconcerningassorted models used in Information retrieval. Here we tend to conjointlylist the research contribution and limitations of Information retrieval modelsfound in various articles.***

***Keywords: Text mining, Information retrieval, NLP, Artificial Intelligence***

## 1. INTRODUCTION

Information retrieval is mostlythought-about as a subfield of computer science that deals with the illustration, storage, and access of data [16],[17]. Information retrieval is worries with the organization and retrieval of information from massive database collections [21],[17].It deals with the search for information and also the illustration, storage and organization of knowledge. Information retrieval is concerned with search processes in which a user needs to identify a subset of information which is relevant for his information need within a large amount of knowledge. The main goal of information retrieval system (IRS) is to "finding relevant information or a document that satisfies user information needs". [17] To attain this goal, IRSs typically follows the process like Indexing, filtering and Searching. The above three are the basic process in information retrieval. Documents area unit described summarized format in indexing. All stop and common words are removed in filtering steps. Searching is performed within the documents based on users need. The paper is organized in the following way section 2 discusses literature review. Section 3 discusses about various available models in information retrieval. Section 4 discusses about merits and demerits of information retrieval. Section 5 research gap and challenges in information retrieval and Section 6 gives conclusion.

## 2. LITERATURE REVIEW

Arpit Deoet al.,[6] presented system; to extract the text from web documents, all mark up language tags are removed. Then stop words and special characters are removed from extracted text for sick solely meaning contents. TF-IDF thought is employed for feature selection. Currently PSO optimization technique is employed for identifying and refining the features set, these selected features are stored in a database that is employed for information retrieval method.

Akram Roshdi et al., [7] represented completely different indexing methods for reducing search space associated different searching techniques for retrieving an information. Thistend to are providing the summary of traditional IR models.

Xiaolu Luet al., [3]conferred QUEST, a technique which will answer complicated questions directly from textual sources on, by computing similarity joins over partial results from totally different documents. This technique is completely unsupervised, avoiding training-data bottlenecks and having the ability to cope with rapidly evolving ad hoc topics and formulation vogue in user questions.

Zong cheng Jiet al., [18] proposed formalizing short text conversation as a pursuit drawback at the primary step, and using progressive information retrieval (IR) techniques to hold out the task. It have tendency to investigate the importance moreover because the limitation of the IR approach. The experiments demonstrate that the retrieval-based model will create the system behave rather "intelligently", once combined with an enormous repository of conversation data from social media.

Achille Souiliaet al., [15] has extracted information from patent documents. He followed four steps such as selected relevant text area. Segmented the paragraph based on matching tags, for parsing XML tags were used. Linguistic phenomena were represented with finite automata. Author considered graph generation as a step before the final step. Author proposed an automated method for extraction of IDM concepts from patent document. Regarding performance analysis author received good precision score, but recall score seems to be very low. This research made him to consider only the partial output. He extends his view that his work has not attained the maturity. The author expressed that he tried to provide assistance for performing mining task easier in patent documents, rather than replacing.

Parul Kalra Bhatia et al., [20] planned a survey done on completely different algorithms that are being worked thus far on PIR systems. Completely different algorithms are being employed to retrieve data within the PIR systems.

Said A. Salloumet al., [9] author did survey in text mining in NLP with Facebook and Twitter, the two media which connects people, where people can communicate, share their views through chats. Most of the data found to be in unstructured format. According to author he decided to consider the result of this research as a base for his future work. Author says that text mining can be classified in to text clustering, classification, association rule mining and trend analysis. Author also says that in nearing future many researches is possible with text mining.

N I Widiastuti [4] in this research conducted in domains of text mining, NLP and deep learning. More than 50 articles used from a various portal of scientific articles. Author says that there exist much research gaps to develop deep learning specifically in text mining and NLP

Anu Bajaj et al.,[2] mentioned the applications of information retrieval with deep learning e.g., web search by reducing the noise and aggregation precise results, trend detection in social media analytics, anomaly detection in music datasets and image retrieval etc.

Ahsan Mahmoodet al.,[7] has implemented Named Entity Recognition (NER) to identify peoples, locations and other entities considered to be one of the fundamental tasks of NLP. Author proposed a knowledge extraction framework to extract named entities from an Urdu book named Hadith. User can extract information through queries from this framework. Author intended to propose a framework and to apply vector space model in his future work.

CharuVirmaniet al., [8] has discussed various NLP approaches necessary for monitoring social network. Author had discussed various open source NLP libraries in this work. Most importantly author revealed many challenges in NLP like parts- of - speech, short context, noisy contents, extracting information about entities. Author had made his contribution by building smart system for analyzing social network information.

SagarGhargeet al., [10] had contributed his research work in detecting malicious tweet using NLP. This system was developed using "Weka" tool and uses SVM machine learning algorithm in classifying the tweet as spam or not spam. In order to evaluate the accuracy of the system, author had chosen 1000 randomly tweet samples, out of which almost 60% were identified to be legitimate tweet and rest were found to be malicious tweet.

Ben He et al.,[23] proposed Okapi BM25, logical model, language model, latent semantic indexing are used for document ranking in IR systems. Term proximity information is found to enhance performance of retrieval function. It's used for XML document retrieval that has a simulating internal structure within the sort of tree. To model this kind of information retrieval system, term proximity is calculated during a distributed environment with totally different resources.

Mourad Sarroutiet al., [11] presented in natural language processing based mostly retrieval system the query within the commencement is fed in to question processing stage. It extracts entity information from the query terms and alternative semantic data. Then the question is classified consistent with the domain, topic and also the entity of focus.

Florian et al., [22] the researchers presented a visual classifier training for text document retrieval with accurate filters. The active learning model reduces the effort of labeling but the support of multi-labeling is not focused.

Shaidah Jusoh[5] had made survey in NLP to identify application, techniques and challenging issues in NLP applications. Author used various papers for his survey from various databases such as Scopus, IEEE explorer and from Google. Where he found information extraction is the one and only most prominent NLP applications. He reviewed question answering system and automated text summarization in his research under NLP applications. He extends his views that ambiguity problem that exists in many languages found to be NLP challenge. Author listed information retrieval, query processing, advanced web search some NLP applications in his research contribution. The future technologies of NLP Chatbot, smart search, text mining, images, audios and videos, visualization of information from text documents like emails, Sms, and reviews and so on were highlighted in this paper.

Mahmoud Othmanet al., [13] proposed his research work in opinion mining using NLP approach. Author proposed a system which classifies opinion in a document or set of documents. Different types of opinion were identified by the system includes categories such as opinionated, comparative, superlative and non-opinionated. Author tested his system with nearly 4000 sentences and evaluated his system with standard metrics such as precision, recall and F-score. Author had decided to focus in multi-language opinion mining in his future research work.

Sweta P. Lendeet al., [14] developed question answering system which extracts information automatically for the asked questions. Author includes documents related to

education using NLP techniques. Author concludes that question answering information retrieval seems to be more complex when compared to other information retrieval system.

## 3. INFORMATION RETRIEVAL MODELS

An Information Retrieval (IR) model specifies the details of the document representation, the query representation and the retrieval functionality. The fundamental IR models can be classified in to Boolean, vector, probabilistic and inference network model [24].

### Boolean Model

In this model, the query is depicted by Boolean expression of terms and therefore the terms area unit connected with Boolean operators. The model may be explained by thinking of a query term as an unambiguous definition of a set of documents. As an example, the query term C merely the set of all documents that's indexed with the term C. The Boolean model permits for the use of operators of Boolean algebra, AND, OR and NOT, for query formulation, however has one major disadvantage: a Boolean system is notable to rank the came back list of documents [26]. Within the Boolean model, a document is related to a set of keywords. Queries are also expressions of keywords separated by AND, OR, or NOT. The retrieval perform during this model treats a document as either relevant or irrelevant [24]. In Figure 1, the retrieved sets area unit visualized by the shaded areas.

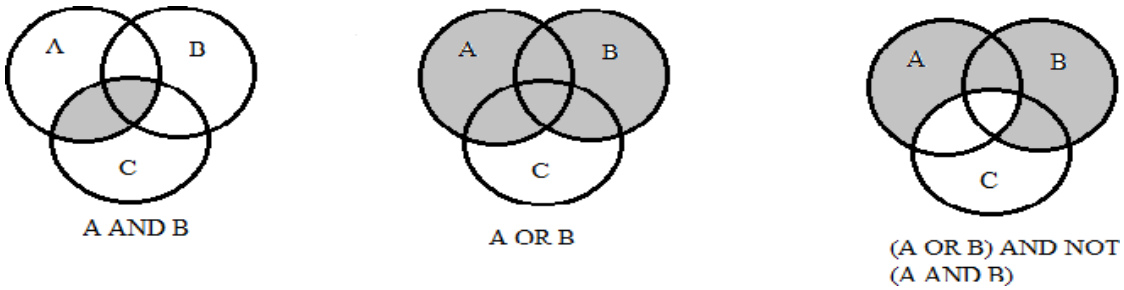


Fig 1. Boolean Combinations of sets visualized as Venn diagrams

### Vector Space Model

In this model, the word and phrases are known as terms and these terms are represented inform of vectors. The vector space model can best be characterized by its attempt to rank documents by the similarity between the query and each document [27].In the Vector Space Model, documents and query are represent as a Vector and the angle between the two vectors are computed using the similarity cosine function. Similarity Cosine function can be defined as:
Where,

$$sim(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \cdot \|q\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w^2{}_{i,j}} \sqrt{\sum_{i=1}^{N} w^2{}_{i,q}}} \quad \text{.....(1)}$$

Documents and queries are represented as vectors

$$d_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$
$$q = (w_{1,q}, w_{2,q}, \dots, w_{t,q}) \quad \text{..... ( 2 )}$$

Vector Space Model have weight scheme referred to as tf - idf weighting. These weights have a term frequency (tf ) factor measuring the frequency of incidence of the terms within the document or query texts associated degree an inverse document frequency (idf) factor measuring the inverse of amount of documents that contain a query or document term [26].

**Probabilistic Model**

It assesses the chance of significance to the query. The documents area unit ordered by decreasing probability of their significance known as probability ranking principle. The foremost vital characteristic of the probabilistic model is it's decide to rank documents by their probability of connexion in a vary given query [28].Documents and queries are depicted by binary vectors ~d and ~q, every vector part indicating whether or not a document attribute or term happens within the document or query, or not. Rather than probabilities, the probabilistic model uses odds O(R), where R suggest
the ‟document is relevant‟ and ¯R means ‟document is not relevant‟ O(R) = P(R)/1 − P(R)[26].

**Inference Network Model**

The documents are modelled using the inference process in the inference networks. The documents are graded in line with the term strength. In this model, document retrieval is modeled as an inference process in an inference network. [29] Most techniques employed by IR systems will be enforced underneath this model. Within the simplest implementation of this model, a document instantiates a term with sure strength, and also the credit from multiple terms is accumulated given a query to figure the equivalent of a numeric score for the document. From associate degree operational perspective, the strength of internal representation of a term for a document will be thought of as the weight of the term within the document, and document ranking within the simplest sort of this model becomes almost like to ranking within the vector space model and also the probabilistic models represented on top of. The strength of internal representation of a term for a document isn't outlined by the model, and any formulation will be used.

**Query based model**

Fig. 2 shows Information Retrieval system architecture [25], [19]. In this figure, the user who needs Information issues a query (**user query**) to the retrieval system through the query operations module. The retrieval module uses the document index to retrieve those documents that contain some query terms. Compute relevance scores for them, and then rank the retrieved documents according to the scores. The ranked documents are then presented to the user. The document collection is also called the text database, which is indexed by the indexer for efficient retrieval [25], [19].The objective of any IR system is to produce a list of relevant documents to the user information need or query provided by the user. The user usually needs relevant documents even if the exact terms he/she used in the provided query were not present in these documents.

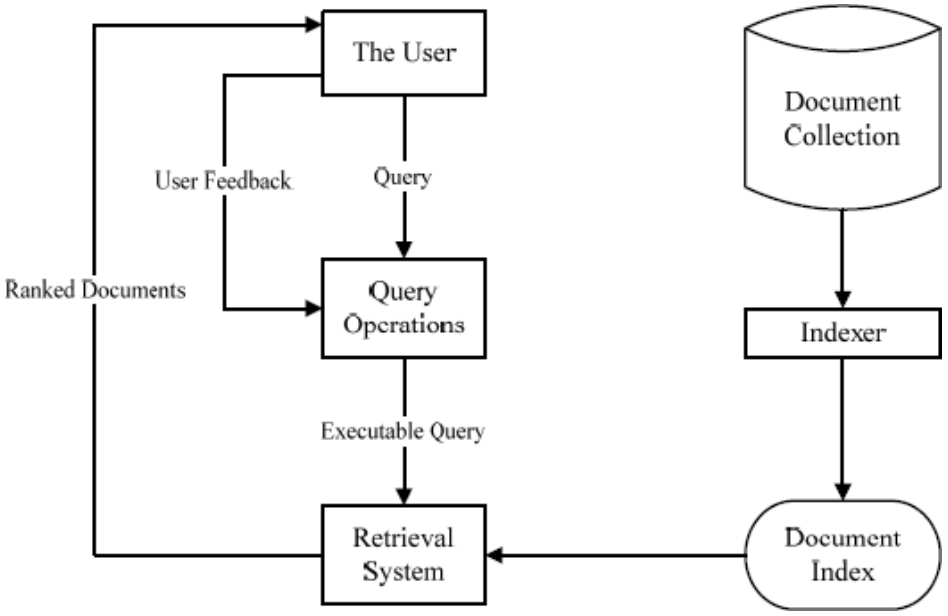


Fig 2. Information Retrieval system architecture

## 4. RESEARCH CONTRIBUTION AND LIMITATIONS OF INFORMATION RETRIEVAL

| Authors | IR Models used | Purpose of Information Retrieval | Method summery and Research Contribution | Limitations |
|---|---|---|---|---|
| B. Piwowarski, 2016 | Probabilistic model | To learn the term weight directly from the occurrences of terms in documents | This has the potential to capture patterns of occurrences that are linked with the importance of a term in a document and a collection. | By not computing term weights for frequent bi or tri-grams. |
| Dwaipayan Roy et al.,2016 | Vector Space model | To develop a similarity metric that makes use of the similarities between the individual embedded word vectors in a | This approach to represent queries and documents as sets of embedded word vectors. A document is diagrammatical as a chance density perform of a | Proposed method gives MAP by up to 5:77%, in comparison to standard text-based language model similarity on |

| | | | | |
|---|---|---|---|---|
| | | document and a query. | combination of Gaussians. | test collection. |
| Chiranjeevi H S et al.,2020 | TF-IDF (Term Frequency-Inverse Document Frequency) Model | A continual Convolution Neural network (RCNN),based mostly text data retrieval system which efficiency retrieves the text documents and data for the user query. | The proposed methodology is not limited to a domain and text document retrieval system. Performance results are extremely promising for organizations that want to extend the customer service effectively. | Nil |
| Anu Bajaj et al., 2020 | Statistical model | The retrieved documents are ranked with the estimate of importance of document for a particular query | Deep learning automatically select raw, heterogonous, high dimensional data, without manual selection. The automatic word detection and topic allocation supported LSTM- RNN for document retrieval. | Nil |
| Anjali Goyal et al., 2017 | SVM-TF-IDF | The information retrieval used for bug report assignment. | Results of the analysis showed a rise of up to 12.8% within the efficiency for the top-5 list size in the information retrieval based technique | An information retrieval based technique which is not considers the time-based expertise computation |
| Kamal Sarkar et al., 2015 | TF-IDF | This paper presents an evaluation and an analysis of some selected information retrieval models For Bengali Monoling ual informati on retrieval task | IR models are often enhanced with adding the power to acknowledge recognize synonyms and/or recognizing similar phrases | Nil |

| | | | | |
|---|---|---|---|---|
| ChahinezBenkoussas et al., 2015 | probabilistic model | This work tackles the problem of book recommendation in the context of INEX (Initiative for the Evaluation of XML retrieval) Social Book Search track | A proposed a novel approach based on Directed Graph of Documents (DGD). It exploits social link structure to enrich the returned document list by traditional retrieval model | It is not explore citation links between scientific documents |

## 5. RESEARCH GAP and CHALLENGES in INFORMATION RETRIEVAL

### 5.1. Research gap in information retrieval with NLP

Some of the research gaps listed was extracted from the existing work
To focus Computation in terms of weights for frequent bi or tri-grams.
To increase MAP of Vector space model up to the MAP of standard text-based language model
To consider time-based expertise in computation of information retrieval
To include citation links while exploring scientific documents
To focus in multi-language opinion mining in his future research work.
To develop deep learning specifically in text mining and NLP

### 5.2 CHALLENGES in INFORMATION RETRIEVAL with NLP

Some of the challenges in NLP are
Finding Parts- of-speech
Working with Short context
Information retrieval from Noisy contents
Extracting information about entities.
Ambiguity problem that exists in many languages found to be popular NLP challenge
Question and answering (query based) information retrieval seems to be more complex and challenging when compared to other information retrieval system.
Most of the social network data are in unstructured format. Handling un-structured data for information retrieval in NLP found to be more complex and Challenging.

## 6. CONCLUSION

Information retrieval in NLP found to be one of the booming research areas. Information retrieval model helps to extract necessary information. This model can be used for different applications. Each model has its own merits and demerits. It provides many new research ideas for researchers. Future research focus of information retrieval can be in using Deep learning specifically in text mining and NLP. Research in text mining can be classified in to text clustering, classification, association rule mining and trend analysis in near future. In future many researches are possible with text mining.

## 7. REFERENCES

[1] GyörgyKovács, Rajkumar Saini, MohamadrezaFaridghasemnia, HamamMokayed, TosinAdewumi, Pedro Alonso, Sumit Rakesh and Marcus Liwicki, "Pedagogical Principles in the Online Teaching of NLP: A Retrospection", *Proceedings of the Fifth Workshop on Teaching NLP*, pages 1–12 June 10–11, 2021.

[2] Anu Bajaj, Tamanna Sharma, Om Prakash Sangwan, "Information Retrieval in Conjunction With Deep Learning", Research Gate, 2020.

[3] Xiaolu Lu, SoumajitPramanik, RishirajSaha Roy, AbdalghaniAbujabal,Yafang Wang, and Gerhard Weikum, "Answering Complex Questions by JoiningMulti-Document Evidence with Quasi Knowledge Graphs",ACM SIGIR Conference (SIGIR'19).ACM, New York, NY, USA,2019.

[4] N I Widiastuti "Deep Learning – Now and Next in Text Mining and Natural Language Processing", IOP Conf. Series: Materials Science and Engineering 407 (2018), pp: 1-6

[5] Shaidah Jusoh, "A Study on NLP Applications And Ambiguity Problems", Journal Of Theoretical And Applied Information Technology, 31st March 2018. Vol.96. No 6, ISSN: 1992-8645, E-ISSN: 1817-3195.

[6] ArpitDeo, JayeshGangrade,"A Survey Paper On Information Retrieval System", International Journal of Advanced Research in Computer Science, 2018,Vol.9(1), pp.778-781.

[7] Ahsan Mahmood, HikmatUllah Khan, Zahoor-ur-Rehman, Wahab Khan, "Query based information retrieval and knowledge extraction using Hadith datasets", 2017, 13th, International Conference on Emerging Technologies (ICET), pp: 1-6.

[8] Charu Virmani, "Extracting Information from Social Network using NLP", International Journal of Computational Intelligence Research, ISSN 0973-1873, Volume 13, Number 4 (2017), pp. 621-630.

[9] Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem, Khaled Shaalan "A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives", Advances in Science, Technology and Engineering Systems Journal, ISSN: 2415-6698, Vol. 2, No. 1, 127-133 (2017)

[10] Sagar Gharge, Mr. ManikChavan , "An Integrated approach for Malicious Tweets detection using NLP",International Conference on Inventive Communication and Computational Technologies, (ICICCT 2017), 978-1-5090-5297-4/17/$31.00 ©2017 IEEE, pp: 435-438

[11] MouradSarrouti , Said Ouatik El Alaoui," A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering", Journal of Biomedical Informatics, Elseiver, 2017, pp. 96–103.

[12] Abhishek Kaushik and SudhanshuNaithani, "A Comprehensive Study of Text Mining Approach", IJCSNS International Journal of Computer Science and Network Security, Vol.16 No.2, February 2016, pp: 69-76.

[13] Mahmoud Othman, Hesham Hassan, Ramadan Moawad, Amira M. Idrees, "Using NLP Approach for Opinion Types Classifier", J*ournal of Computers,* Volume 11, Number 5, September 2016, pp: 400-410.

[14] Sweta P. LendeDr.M.M. Raghuwanshi **,** "Question Answering System on Education Acts Using NLP Techniques", IEEE Sponsored World Conference on Futuristic

Trends in Research and Innovation for Social Welfare (WCFTR'16), 978-1-4673-9214-3/16/$31.00 © 2016, pp:1-6.
[15] Achille Souilia, Denis Cavalluccia , François Rousselotb, "Natural Language Processing (NLP) - A solution for knowledge extraction from patent unstructured data", Procedia Engineering 131 ( 2015 ) 635 – 643.
[16] M.FrançoisSy, S.Ranwez, J.Montmain,"User centered and ontology based information Retrieval system for life sciences", BMC Bioinformatics, 2015.
[17] AkramRoshdi, AkramRoohparvar, "Review: Information Retrieval Techniques and Applications", International Journal of Computer Networks and Communications Security, 2015, Vol: 3, No: 9,pp. 373–377.
[18] Zongcheng Ji, Zhengdong Lu, Hang Li, "An Information Retrieval Approach to Short Text Conversation",arXiv,2014
[19] Yogesh Gupta, Ashish Saini, A.K. Saxena, "A Review on Important Aspects of Information Retrieval", International Journal of Computer and Information Engineering, 2013, Vol:7, No:12.
[20] ParulKalra Bhatia, Tanya Mathur, TanayaGupta,"Survey Paper on Information Retrieval Algorithms and Personalized Information Retrieval Concept", International Journal of Computer Applications,2013,Vol. 66(6), pp.14-18.
[21] R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", IJCER, sep 2012, Vol. 2 Issue. 5, PP: 1443-1444.
[22] Florian, H., Koch, S., Bosch, H., &Ertl, T." Visual classifier training for text document retrieval",*IEEE Transactions on Visualization and Computer Graphics,* 2012,Vol.*18*(12), pp. 2839–2848.
[23] Ben He, Jimmy XiangjiHuang , Xiao feng Zhou," Modeling term proximity for probabilistic information retrieval models", Information Sciences, 2011, pp.3017–3031
[24] Anwar A. Alhenshiri, "Web Information Retrieval and Search Engines Techniques", Al-Satil journal, 2010, PP: 55-92.
[25] Liu B," Web Data Mining: Exploring Hyperlinks, Contents andUsage Data", Springer-Verlag, Berlin Heidelberg, 2007.
[26] D.Hiemstra,P. de Vries, "Relating the new language models of information retrieval to the traditional retrieval models", published as CTIT technical report TR-CTIT-00-09, 2000.
[27] H. Turtle, "Inference Networks for Document Retrieval". Ph.D. thesis, Department of Computer Science, University of Massachusetts, Amherst, MA 01003. Available as COINS Technical Report 90-92, 1990.
[28] G. Salton and M.J. Mc Gill, "Introduction to Modern Information Retrieval". McGraw-Hill , 1983.
[29] C. J. van Rijsbergen. "Information Retrieval. Butterworths", London, 1979