ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Information Retrieval: A Study of Various Models and Methods

Gurpreet Kaur
Assistant Professor
Department of Computer Science
Guru Nanak Khalsa Girls College,
Baba Sang Dhesian Phagwara, Punjab

Abstract: *Information Retrieval is become a research area in the field of Computer Science. Information Retrieval is the process of tracing and recovery of specific information from Database.It is a continuous process during which we consider, reconsider and refine the research problems by using different retrieval techniques. It is crucial to documentation and organization of knowledge. This paper provides aComprehensive study of major Information Retrieval Models and Methods and their Applications.*
*Keywords:*Information Retrieval (IR), Indexing, IR mode,Searching, Vector Space Model (VSM).

1.INTRODUCTION

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). It is the task of representing, storing, organizing, and offering access to information items. IR is different from data retrieval, which is about finding precise data in databases with a given structure. Information retrieval is generally considered as a subfield of computer science that deals with the representation, storage, and access of information [1]. Information retrieval is concerned with the organization and retrieval of information from large database collections [2]. Information Retrieval (IR) is the process by which a collection of data is represented, stored, and searched for the purpose of knowledge discovery as a response to a user request (query) [3]. Information retrieval is a *problem-oriented* discipline, concerned with the problem of the effective and efficient transfer of desired information between human generator and human user In other words:

• The indexing and retrieval of textual documents.
• Concerned firstly with retrieving *relevant* documents to a query.
• Concerned secondly with retrieving from *large* sets of documents *efficiently*.

An Example:

- Suppose there is a store of documents and a person (user of the store) formulates a question (request or query) to which the answer is a set of documents satisfying the information need expressed by his question.
- Solution: User can read all the documents in the store retain the relevant documents and discard all the others – *Perfect Retrieval… NOT POSSIBLE!!!*
- Alternative: Use a High Speed Computer to *read* entire document collection and extract the *relevant* documents.
- **Goal** = find documents *relevant* to an information need from a large document set

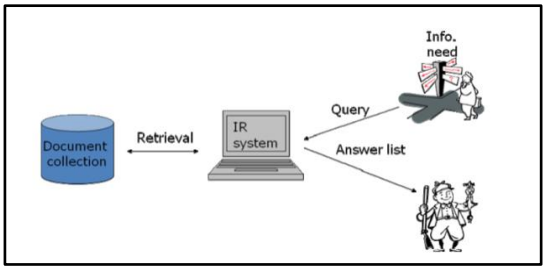

Fig: 1. Basic Structure of IR

IR Process

IR process involves various stages initiate with representing data and ending with returning relevant information to the user. Intermediate stage includes filtering, searching, matching and ranking operations. The main goal of information retrieval system (IRS) is to "finding relevant information or a document that satisfies user information needs". The Figure2.sketches the outline of information process.

Here, the user issues a query *q* from the front-end application (accessible via,e.g., a Web browser); *q* is processed by a *query interaction* module that transforms it into a "machine-readable" query *q*' to be fed into the core of the system, a *search and query analysis* module. This is the part of the IR system having access to the *content management* module directly linked with the back-end information source (e.g., a database). Once a set of results *r* is made ready by the search module, it is returned to the user via the result interaction module; optionally, the result is modified (into *r* ') or updated until the user is completely satisfied.

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

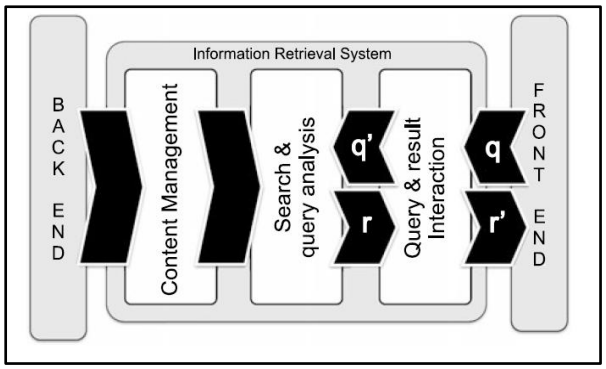

Fig:2. A high-level view of the IR process

2.IR MODELS

An IR model specifies the details of thedocument representation, the query representationand the retrieval functionality .The fundamental IR models can be classified intoBoolean, vector, probabilistic and inference network model. The rest of this sectionbriefly describes these models.

2.1Boolean Model

The Boolean model is the first model ofinformation retrieval and probably also the mostcriticised model. The model can beexplained by thinking of a query term as anunambiguous definition of a set of documents. Forinstance, the query term economic simply definesthe set of all documents that are indexed with theterm economic. Using the operators of GeorgeBoole's mathematical logic, query terms and theircorresponding sets of documents can be combined to form new sets of documents. The Boolean model allows for the use of operators of Boolean algebra, AND, OR and NOT, for query formulation, but has one major disadvantage: a Boolean system is notable to rank the returned list of documents . Inthe Boolean model, a document is associated with aset of keywords. Queries are also expressions ofkeywords separated by AND, OR, or NOT/BUT.The retrieval function in this model treats adocument as either relevant or irrelevant. In **Figure 3**, the retrieved sets are visualised by the shaded areas.

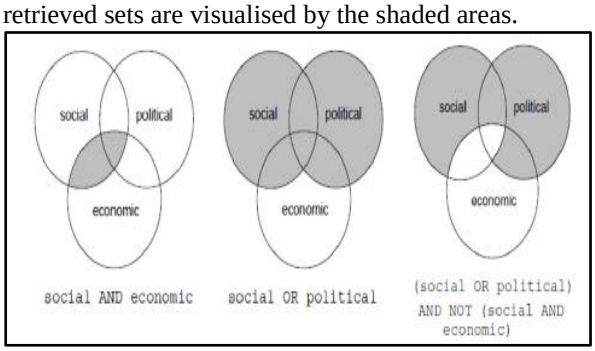

Fig:3Boolean combinations of sets visualised as Venndiagrams

2.2 Vector Space Model

Gerard Salton and his colleagues suggested amodel based on Luhn's similarity criterion that has a stronger theoretical motivation (Salton andMcGill 1983). They considered the indexrepresentations and the query as vectors embedded in a high dimensional Euclidean space, where each term is assigned a separate dimension. The vectorspace model can best be characterized by itsattempt to rank documents by the similaritybetween the query and each document .In theVector Space Model(VSM), documents and query are represent as a Vector and the angle between the two vectors are computed using the similaritycosine function. Similarity Cosine function can be defined as:

Where,

$$\mathrm{sim}(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \, \|\mathbf{q}\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

Documents and queries are represented as vectors.

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

$$q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$

Vector Space Model have been introduce termweight scheme known as if-idf weighting. Theseweights have a term frequency (tf) factormeasuring the frequency of occurrence of the terms in the document or query texts and an inversedocument frequency (idf) factor measuring theinverse of the number of documents that contain a query or document term .

2.3 Probabilistic Model

Whereas Maron and Kuhns introduced rankingby the probability of relevance, it was StephenRobertson who turned the idea into a principle. He formulated the probability ranking principle, which he attributed to William Cooper, as follows(Robertson 1977). The most important characteristicof the probabilistic model is its attempt torank documents by their probability of relevance given a query. Documents and queries arerepresented by binary vectors ~d and ~q, each vector element indicating whether a documentattribute or term occurs in the document or query,or not. Instead of probabilities, the probabilisticmodel uses odds O(R), where O(R) = P(R)/1 P(R), R means ‟document is relevant‟ and ‘Rmeans ‟document is not relevant‟.

2.4 Inference Network Model

In this model, document retrieval is modeled asan inference process in an inference network. Most techniques used by IR systems can beimplemented under this model. In the simplestimplementation of this model, a document instantiatesa term with a certain strength, and the creditfrom multiple terms is accumulated given a queryto compute the equivalent of a numeric score forthe document. From an operational perspective, thestrength of instantiation of a term for a documentcan be considered as the *weight* of the term in thedocument, and document ranking in the simplestform of this model becomes similar to ranking inthe vector space model and the probabilistic models described above. The strength of instantiation of a term for a document is not defined by the model,and any formulation can be used.

978-93-85670-72-5 © 2016 (RTCSIT)

3 INDEXING TECHNIQUES

There are several popular information retrievalindexing techniques, including inverted indices and signature files.

3.1 Signature File

In signature file method each document yields abit string („signature") using hashing on its wordsand superimposed coding. The resulting documentsignatures are stored sequentially in a separate filecalled signature file, which is much smaller than theoriginal file, and can be searched much faster.

3.2 Inversion Indices

Each document can be represented by a list ofkeywords which describe the contents of thedocument for retrieval purposes [6]. Fast retrievalcan be achieved if we invert on those keywords.The keywords are stored, eg alphabetically; in theindex file for each keyword we maintain a list ofpointers to the qualifying documents in the postings file. This method is followed by almost all thecommercial systems.

4 SEARCHING TECHNIQUES

There are various searching algorithms, includinglinear search, binary search, brute force search etc.
some general searching algorithms are describedbelow:
1) In linear search algorithm is a method offinding a particular element or keyword fromlist or array that checks every element in list,one at a time and in sequence. Linear searchis a simplest search algorithm. One of themost important drawbacks of linear search isslow searching speed in ordered list. Thissearch is also known as sequential search.
2) Brute force search is a very general problemsolvingtechnique that consists ofsystematically enumerating all possiblecandidates for the solution and checkingwhether each candidate satisfies theproblem's statement. Brute force algorithm issimple to implement and it will always find asolution if it exist.
3) Binary search algorithm, finds specifiedposition of the element by using the keyvalue with in a sorted array. In each step, thealgorithm compares the search key valuewith the key value of the middle element ofthe array. If the keys match, then a matchingelement has been found and its index, orposition, is returned. Otherwise, if the searchkey is less than the middle element's key,then the algorithm repeats its action on thesub-array to the left of the middle elementor, if the search key is greater, on the subarrayto the right.
If the remaining array to be searched is empty,then the key cannot be found in the array and aspecial "not found" indication is returned.

5. TYPES OF IR SYSTEMS

There are three types of IR systems,which are explained below.

5.1 Textual Retrieval

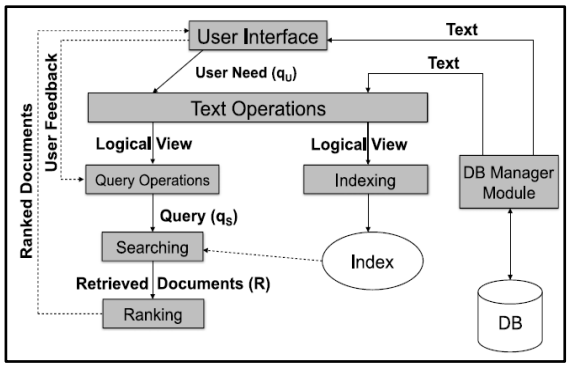

Fig: 4Architecture of a textual IR system.

Textual operations translate the user's need into a logical query and create a logical view of documents.The most widespread applications of IR are the ones dealing with textual data.As textual IR deals with document sources and questions, both expressed in naturallanguage, a number of textual operations take place "on top" of the classic retrievalsteps. Figure4sketches the processing of textual queries typically performed byan IR engine:
1. The user need is specified via the user interface, in the form of a textual *queryqU*(typically made of keywords).
2. The query *qU*is parsed and transformed by a set of textual operations; the sameoperations have been previously applied to the contents indexed by the IR system; this step yields a refined query *q'U*.
3. Query operations further transform the preprocessed query into a system-levelrepresentation, *qS*.Textual IR exploits a sequence of text operations that translate the user's needand the original content of textual documents into a logical representation moreamenable to indexing and querying.

5.2 Document Retrieval

In Document Retrieval, some processes take place dynamically when the user inputs their query, while other processes take place off-line in advance and in batch mode and do not involve individual users. These static processes are run on the documents that will be made available in the retrieval system. These will be explained first. Then, the two dynamic processes, Query Processing and Matching, will be presented. **Figure 5.**provides a simple, but clear view of the relationship between these three processes.

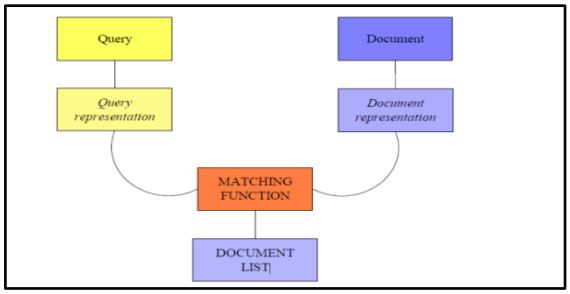

Fig: 5. Basic Components of a Document Retrieval System

978-93-85670-72-5 © 2016 (RTCSIT)

Document Processing: The first two steps in the processing of documents are somewhatmundane, but necessary, and can be considered as batch pre-processing. These are:

1. Normalize document stream to a predefined format, whereby multiple externalformats (e.g. newsfeeds, web pages, word processed documents) are standardized intoa single consistent format. This is an essential step (much akin to data clean-up indata mining) as all downstream processes rely on receiving a common format theycan recognize and process. Preprocessing is particularly vital for systems with morecomplex processing than simple 'characters between white spaces' indexing.
2. Break document stream into desired retrievable units, whether this is web page,chapter, full document, paragraph, etc. The pointers stored in the inverted file are towhatever unit size has been pre-determined. Therefore, document retrieval could infact be paragraph retrieval, if the indexable unit was determined at this stage to be theparagraph.
3. Identify potential indexable elements in documents. This is a key decision pointthat dramatically affects the nature and quality of the retrieval performance. First, theimportant definition needs to be made as to what is a term. Is it any string of alphanumericcharacters between blank spaces or punctuation? If so, are noncompositionalphrases or multi-word proper names, or inter-word symbols such ashyphens or apostrophes treated differently (e.g. are "small business men" and "smallbusinessmen" the same)? At this stage, the system requires a set of rules to beexecuted which control what actions are taken by the 'tokenizer' – the algorithmwhich recognizes 'indexable terms'. IR systems vary as to which of these processesthey perform, but the most frequently used processes are:

a. Delete stop words via an algorithm that filters the document's potentialindexable elements against a Stop Word list to eliminate terms that aredeemed to be insignificant in determining a document's relevance to a user'srequest. The original objective in using stop words was to save systemresources by eliminating those terms that have little value for retrievalperformance.

b. Stem terms by removing suffixes. In this morphological step, some IRsystems do just inflectional ('weak') stemming which only changes thesubclass within a part-of-speech category, i.e. past tense to present tense,while others also do derivational ('strong') stemming which removes suffixes,sometimes recursively, that may actually change the part of speech of a word.Use of stemming will result in fewer entries in an index, each of which islikely to have higher frequency counts than if all morphological variants andtheir counts are used.

c. Bracket noun phrases, usually by means of regular expressions which definethe part-of-speech patterns which comprise a noun phrase (e.g. <ADJ NN> or<NN NN>). This is a step that can negatively affect recall of retrieval resultsby either excluding documents when the phrasal expression in the query is notexactly the same as the index entry of a document, or positively affectprecision by retrieving only documents that include the terms in the desiredphrasal expression.

4. Produce an inverted file containing a sorted array of all indexable terms (with termdefined as referring to either a word or a phrase), along with the unique identificationnumber of each document in the collection in which the term occurs, a link to each ofthese documents, weights for each term as determined by the IR model beingimplemented in the system andoptionally, the within-document location of the term. More sophisticated systems mayinclude further information in the inverted file, such as named entity category forProper Names (i.e. PERSON, ORGANIZATION, GEO-LOCATION, etc) but themost common features are simply term, document ID, and weight.Query Processing: The system's internal representation of the user's question / searchterms is typically referred to as the query. Most of the same processes that are run on thedocuments are also run to produce the query, but there are some unique processes as well.As distinct from document processing, all of the query processing is done in real time,while the user awaits their documents. These are:

1. Recognize query terms vs. special operators, such as "I need information about…"which do not convey the topic of the user's information need and will not be includedin the query representation.
2. Tokenize query terms, a process that requires similar decisions as were described onthe document processing side – that is stop word deletion, stemming, and phraserecognition.
3. Create query representation, which typically follows stop word removal andstemming, and which may also include insertion of logical operators between /amongst terms requiring cooccurrence or simple presence of only one of thearguments.
4. Expand query terms to include variant terms that refer to or relate to the sameconcept. Queryexpansion relieves the user of needing to generate all conceptual variants of theirsearch terms and is likely to improve recall, but may reduce precision when erroneoussenses of the newly introduced terms retrieve irrelevant documents. The longer aquery is, the less likelihood that erroneous senses of expanded terms will have anegative impact, but also the less likely that expansion will contribute much to theretrieval results.
5. Compute query term weights. This step is less commonly included in DocumentRetrieval systems, mainly because it is difficult both for users to know how to assignweights to query terms in a way that improves retrieval results, or for automaticweighting, since queries are frequently so short as to give little evidence of therelative importance of query terms as most terms only occur once in a single query.The process description below may be easier to follow if youconceive of both the query and the documents as vectors of terms, with frequencyinformation or weights for each term in the vector.

1. Search inverted file for documents that contain terms in the query. This is typicallydone using a standard binary search. Each document that contains any of the queryterms becomes a candidate for retrieval.
2. Compute similarity score between query and each candidate document using thealgorithm prescribed by one of the four Document Retrieval models being used. Thisscore is referred to as the Similarity Coefficient.
3. Rank order the documents in decreasing order based on the scores assigned themby the scoring algorithm.
4. Provide list of perceived relevant documents to user ranked by similarity scorebetween query and document. Systems that utilize other sources of evidence of valueof a document to the query, such as number of links from the page/document to

978-93-85670-72-5 © 2016 (RTCSIT)
CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

orfrom other pages/documents, would integrate this information and produce apotentially different ranked list.

5. Allow for query modification by the user if user-based relevance feedback isprovided by the system. If so, typically, the user marks the documents they findrelevant, either based on just the title and brief description shown them in the initiallist or by actually reviewing the full document, which they can link to from the resultspage.

6. Perform relevance feedback based on user's input. The algorithm for user-basedrelevance feedback is typically the same as that for automatic relevance feedback asdescribed in Step 3 above. The system then re-runs the search with the revised, andhopefully improved, query and produces a revised ranked list of documents. Therelevance feedback loop is iterative and can be performed as many times as the userwants.

5.3 Image Retrieval Systems

Content-based image retrieval, uses the visual contents of an image such as *color*,*shape*, *texture*, and *spatial layout* to represent and index the image.

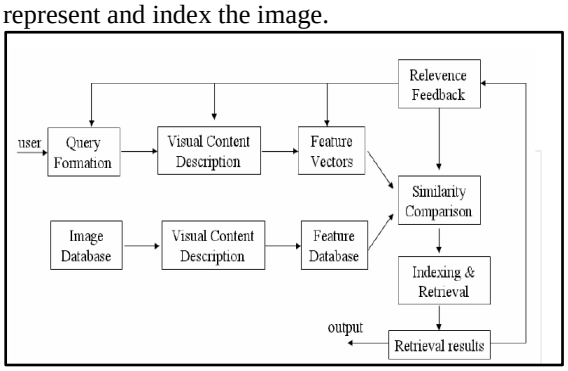

Fig: 6 Diagram for content-based image retrieval system

In typical content-based image retrieval systems (Figure6), the visual contents of the images in the database are extracted and described by multi-dimensional feature vectors. The feature vectors of the images in the database form a feature database. To retrieve images, users provide the retrieval system with example images or sketched figures.

The system then changes these examples into its internal representation of feature vectors. The similarities /distances between the feature vectors of the query example or sketch and those of the images in the database are then calculated and retrieval is performed with the aid of an indexing scheme. The indexing scheme provides an efficient way to search for the image database. Recent retrieval systems have incorporated users' relevance feedback to modify the retrieval process in order to

generate perceptually and semantically more meaningful retrieval results.

6.APPLICATIONS

Information retrieval is used today in many applications. It is used to search for documents,content thereof, document metadata within traditional relational databases or internet documents more conveniently and decrease work to access information. Retrieved documents should be relevant to a user's information need. Obvious examples include search engines as Google,Yahoo or Microsoft Live Search.

6.1 Digital Library

A digital library is a library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers. The digital content may be stored locally, or accessed remotely via computer networks. A digital library is a type of information retrieval system.Many academic libraries are actively involved in building institutional repositories of the institution's books, papers, theses, and other works which can be digitized or were 'born digital'. Many of these repositories are made available to the general public with few restrictions, in accordance with the goals of open access, in contrast to the publication of research in commercial journals, where the publishers often limit access rights. Institutional, truly free, and corporate repositories are sometimes referred to as digital libraries.

6.2 Recommender systems

Recommender systems or recommendation engines form or work from a specific type of information filtering system technique that attempts to recommend information items (films, television, video on demand, music, books, news, images, web pages, etc) that are likely to be of interest to the user. Typically, a recommender system compares a user profile to some reference characteristics, and seeks to predict the 'rating' that a user would give to an item they had not yet considered. These characteristics may be from the information item (the content-based approach) or the user's social environment (the collaborative filtering approach).Collaborative filtering is concerned with making recommendation about informationitems (movies, music, books, news, web pages) to users.

6.3 Search Engines

A search engine is one of the most the practical application of information retrieval techniques to large scale text collections. Web search engines are best-known examples, but many others searches exist, like: Desktop search, Enterprisesearch, Federatedsearch, Mobilesearch, Social search.

A web search engine is designed to search for information on the World Wide Web. The search results are usually presented in a list of results and are commonly called *hits*. The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike Web directories, which are maintained by human editors, search engines operate algorithmically or are a mixture of algorithmic and human input. Reliability of information is a pre-requisite to get most from research information found onto the web. A frequently encountered issue is that search terms are ambiguous and thus documents from a different non-relevant context are retrieved or you may not know which terms describe your problem properly, especially if you are a non-expert user in this particular domain.The novel idea of relevance feedback allows users to rate retrieved documents as relevant or less relevant and thus help other users to find

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

documents more quickly. These ideas where adopted from image retrieval. Images are hard to describe using words.

6.4 Media search

An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. Most traditional and common methods of image retrieval utilize some method of adding metadata such as captioning, keywords, or descriptions to the images so that retrieval can be performed over the annotation words. Manual image annotation is time-consuming, laborious and expensive; to address this, there has been a large amount of research done on automatic image annotation. Additionally,the increase in social web applications and the semantic.

7. CONCLUSION

Information retrieval is a process of searching and retrieving the knowledge based information from collection of documents. This study dealt with the basics of the information retrieval. In this paper we have also discussed about the different indexing techniques and searching techniques. This paper also includes the area of IR applications.

REFERENCES

[1] M.FrançoisSy, S.Ranwez, J.Montmain,"Usercentered and ontology based informationRetrieval system for life sciences", BMCBioinformatics,2105.

[2] R. Sagayam, S.Srinivasan, S. Roshni, "ASurvey of Text Mining: Retrieval, Extractionand Indexing Techniques", IJCER, sep 2012,Vol. 2 Issue. 5, , PP: 1443-1444,.

[3]Clarke, C., G. Cormack, and F. Burkowski (1995).Algebra for structuredtext search and a framework for its implementation.The Computer Journal 38 (1), 43{56.

[4]Information Retrieval Data Structures & Algorithms - William B. Frakes and Ricardo Baeza-Yates

[5]Introduction to Information Retrieval – Christopher D. Manning, PrabhakarRaghavan ,HinrichSchutze

[6]Data Mining: Concepts and Techniques - Jiawei Han &MichelineKamber

[7] Algorithms for Information Retrieval – Introduction, Lab module 1.

[8] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval", ACM Press, ISBN: 0-201-39829-X.

[9] S.E. Robertson and K. Sparck Jones.Relevance weighting of search terms. Journal of the American Society for Information Science,27:129–146, 1976.

[10] G. Salton and M.J. McGill, editors.Introduction to Modern Information Retrieval. McGraw-Hill 1983,

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India