www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 7 July 2015, Page No. 13578-13580

Information Retrieval

Mukta Airon
Department of Computer Science
mukta.mittal2006@gmail.com

ABSTRACT

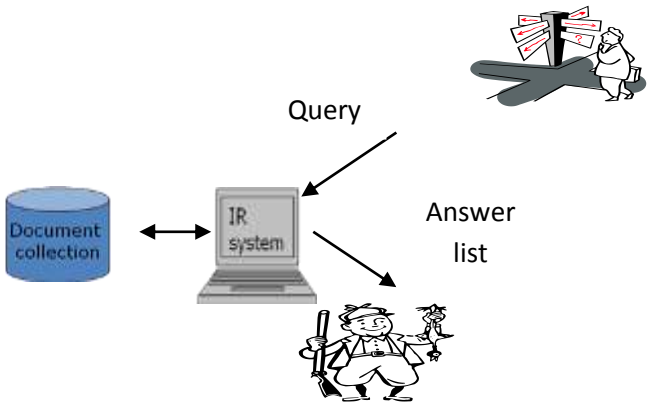
The field information Retrieval deals with representing, storage and access to information items.IR is the most usual way of information access, mostly due to the increasing widespread of world wide web(WWW).Information Retrieval mainly deals with retrieval of unstructured data, especially textual documents, in response to a query or topic statement, which may itself be unstructured. In this paper we present the introduction to the information Retrieval and focuses on properties, key techniques and term weighting factor of information retrieval.

Keywords-Information Retrieval, Text Extraction, Term frequency, inverse document frequency and term weighting.

1. INTRODUCTION

Given the increasing amount of information that is available today, there is a clear need for information Retrieval (IR) systems that can process this information in an efficient and effective way.Efficient processing means minimising the amount of time and space required to process data, whereas Effective means identifying accurately which information is relevant to the user and which is not. The concept of Information in this context strongly bound to a user's request.IR system takes user's information need, normally in the form of query and they must return information related to the query, in the form of documents.IR systems present the information as a ranking of documents.IR deals with the problem of finding and presenting the documents of unstructured data that satisfy information need from within the collection of documents. Unstructured refers to the data which is unambiguous. The term document refers to the information presented to the user. It can represent abstract, articles, WebPages, book chapters, emails, sentences and so on. The term document is not restricted to textual documents, as users may be interested in retrieving and accessing multimedia data like video, audio or images. Collection may refer to a repository of documents from which information is retrieved. A user information need also referred to a query, must be translated in order for an IR systems to process it and retrieve information relevant to its topic. This translation is usually made by extracting a set of keywords that summarises the description of information need. At last presentation of documents in such a way that facilitates the user to find the documents that he/ she is interested in. The most difficult part of the retrieval process is deciding which documents are related to or satisfy a certain query. Documents should be ranked in decreasing order of relevance. An IR system achieves its maximum effectiveness when the relevant documents with respect to the user's query are ranked higher, whereas the non-relevant are ranked lower. Relevance depends not only on the query and collections but also on the context e.g. user's personal needs, preferences, knowledge expertise, language etc. Hence a given document retrieve by a query is may be relevant to the user on one day but not on another. Given document may be relevant to one user not to another even though they both used the same query.

INFORMATION RETRIEVAL SYSTEM



DATA STRUCTURE- most IR system based on Inverted list data structure. This enables fast access to a list of documents that contain term along with other information e.g. weight of the term in each document, relative positon of the term in each document.

Inverted list may be stored as :-

$t_i \rightarrow < d_a, \ldots\ldots >, < d_b \ldots\ldots >, < d_n \ldots.. >$

Which shows that term i is contained in doc. $d_a, d_b, \ldots, d_n$ and stores any other information.

Models of information retrieval (Boolean Model, Vector-space model, probabilistic model) are implemented by inverted lists.

Inverted list explore the fact that most IR system are only interested in storing a small number of documents that contain same query term. This allows system to only score documents that have a non-zero numeric score. All documents are indexed by the term they contain. The process of generating, Building and storing documents representation is called Indexing and restoring inverted file are called the inverted index.

PROPERTIES Two Properties that have been accepted by IR community for measurement of search Effectiveness are Recall and Precision.

Recall: The Proportion of relevant documents retrieve by the system.

Precision: The proportion of retrieved documents that are relevant.

Good IR System should have a high Recall should retrieve as many relevant documents as possible. It should have high precision means may retrieve few non- relevant documents.

KEY TECHNIQUES The most critical piece of information needed for document ranking are:

1) In all models is a term's weight in doc.
2) Another technique that is to be effective in improving document ranking is query modification via relevance feedback.

TERM WEIGHTING

In all models three main factors came into play in find out the term weight formulation are:

Term frequency (tf), Inverse Document Frequency (idf) and tf-idf.

tf-idf short for term frequency- inverse document frequency is a numerical statistic that reflects how important a word to a document in a collection or corpus.

i) It is used as a weighting factor in information retrieval and text mining.
ii) tf-idf values increase proportionally to the no of times a word appear in a document.
iii) tf-idf weighting scheme are used by search engines as a central tool in scoring and ranking a doc's relevance given a user query.
iv) tf-idf can be successfully used for stop words filtering .
v) Simplest ranking function is computed by summing tf-idf for each query term.

TERM FREQUENCY (tf) – the number of times a term occurs in a document is called term frequency.

e.g. "the brown cow " A simple way to start out is by eliminating documents that do not contain all three words "the", "brown " ,"cow". The term "the" is not a good keyword to distinguish relevant and non-relevant documents and terms, unlike the less common words "brown "and "cow". Hence an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

Weight of a term that occurs in a document is simply proportional to the term frequency.

Term frequency tf(t.d)= f(t,d)

Boolean "frequencies":tf(t,d)=1 if t occurs in d and 0 otherwise;

Logarithmically scaled frequency :tf(t,d)=1+logf(t,d),or zero if f(t,d)is zero.

INVERSE DOCUMENT FREQUENCY: - is a measure of how much information the word provides that is whether the term is common or rare across all documents. It is the logarithmically scaled fraction of documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of the quotient.

Idf(t,d)

tf-idf is the product of two statistics, term frequency and inverse document frequency. Then tf-idf is calculated as

tf-idf (t, d, D) = tf(t,D)×idf(t,D)

We can calculate tf-idf as:

tf-idf= $f_{t,d} \times \log N/n_t$

So in this way we can calculate the weighting factor in information Retrieval and text mining.

RELEVANCE FEEDBACK is used to improve query. It is a feature of information retrieval. The idea behind relevance feedback is to take the results that are initially returned from a given query and to use information about whether or not those result are relevant to perform a new query.

FUTURE SCOPE Information retrieval research has reached a point where it is appropriate to assess progress and to define a research agenda for the next five to ten years. Major IR challenges are retrieval models, cross-lingual retrieval, Web search, user modelling, filtering, topic detection and tracking, classification, summarization, question-answering, metasearch, distributed retrieval, multimedia retrieval, information extraction for future scope. Contextual retrieval and global information retrieval access were identified as particularly long-term challenges.

REFERNCES

1) S.D.Manning "An introduction to information Retrieval"in online edition 2009 UP.

2) A.Baeza and A.Riberio-"Modern information Retrieval "ACM 1999.ISBN 0-2-1-39829.

3) Ngoc Anh and A.moffat "Effectiveness and Efficiency of web retrieval ".n SIGIR '02' proceedings of 25th annual international conference on research and development in information retrieval.

4) By Kit Marlow CIS650 02/26/03"Information Retrieval Models".

5)Dagobert Soergel " Important problems in Information Retrieval " college of library and information services in August 1989.

6) By jian-yun Nie University of Montreal Canada "Information of IR system ".

7)"Challenges in Information and Language modelling "by James allan,Jay Aslam in year 2002,university of Massachusetts Amherst.

8)AI-Maskari , M.Sanderson and P.clough,"the relationship between IR effectiveness measure and user satisfaction ".

9) J. articles,S.Sekine and J.Gonzalo,"Web people search:Result of the first evaluation and plan for the second," in proceeding of the 17th international conferences on world wide web.

10) M.Baillie , L.Azzopardi and I.Ruthven "a retrieval evaluation methodology for incomplete relevance assessment " in Advances in information retrieval lecture notes in computer science,vol 4425.