INFORMATION RETRIEVAL USING PAGE RELEVANCY

Jaswinder Singh¹, and Jyoti Godara²

¹Department of Computer Science & Engineering, Guru Jambheshwar University Of Science & Technology, Haryana, INDIA, E-mail: jaswinder_singh_2k@rediffmail.com,
²Department of Computer Science & Engineering, Guru Jambheshwar University Of Science & Technology, Haryana, INDIA, E-mail: jyotigodara12@gmail.com

ABSTRACT

Retrieval of relevant documents from a large collection of web contents is done through Information Retrieval System. Relevancy of a page can be calculated with the use of similarity measure. Thus similarity measures ranks the pages according to the relevancy of each page and helps the information retrieval system to present to the users the more relevant pages earlier in response to a query. In this paper, a general frame of Information Retrieval system and search engines are discussed. Various Information retrieval models and types of similarity measures are also presented.

Keywords: Information retrieval, page relevancy, similarity measure, web crawlers, search engine, crawler.

1. INTRODUCTION

Recent developments on the computer and networking technologies have made the Internet to be the most popular and the largest information source over the world. It was found that about a decade ago, the size of the Web was doubled every year as the Web contained more than 350 million pages and 600 Gigabytes of information on these pages were updated every month [1]. Thus users are facing information overloaded. Search engines are used by most people to find whatever information on the Web. Given a few search keywords, most search engines today will retrieve more than a few thousand Web pages. The problem now is that, to find what users require, they need to scan pages after pages, manually and time consuming or often give up without getting the needed information. There are several approaches to address the problem. The currently most popular method is page ranking in which the search results are ordered and presented to the users, the most relevant pages first. [2]

1.1. Search Engine

Search engine is a software program that searches documents for specified keywords and returns a list of the documents where the keywords are found. Web search engines work by storing information about many web pages. These pages are retrieved by a Web Crawler. The contents of each page are then analyzed to determine how it should be indexed. Data about web pages are stored in an index database for use in later queries. The purpose of indexing is to allow information to be found quickly. When a user enters a query, the search engine results the best-matching web pages according to its criteria. The index is built from the information stored with the data and the method by which the information is indexed. Most search engines employ methods to rank the results to provide the best results first. The usefulness of a search engine depends on the relevancy of the result set it gives back.

1.2. Information Retrieval System

Information retrieval (IR) is the area of study concerned with searching for documents, for metadata about documents, as well as that of searching structured storage, relational database, and the World Wide Web. An information retrieval process begins when a user enters a query into the system. Here, a query does not uniquely identify a single object in the collection; instead, several objects may match the query, perhaps with different degrees of relevancy.

A document-based IR system (Fig. 1) typically consists of three main subsystems: Document Database, Query Subsystem, and Matching Mechanism. Documentary database stores the documents and their representations. Document collection consists of many documents containing information about various subjects. Document contents are transformed into a document representation in such a way that matching these with queries is easy. Typically representation proceeds by extracting keywords that are considered as content identifiers and organizing them into a given format. Queries transform the user’s information requirement into such a form that is suitable for the matching process. Query formatting depends on the underlying model of information retrieval used. Matching Mechanism matches a user’s queries with the
document representations and retrieves the documents that are most likely to be relevant to the user. A matching algorithm addresses two issues: 1. How to decide how well the document matches a user’s information request. If a document term does not match search terms then a relevant document may not be retrieved. 2. How to decide the order in which the documents are to be presented to the user. Typically the matching algorithms calculate a matching number for each document and documents are retrieved in the decreasing order of this number. The user rates documents shown as either relevant or non-relevant to his/her information need. The basic problem facing any IR system is how to retrieve only the relevant documents for the user’s information need, while not retrieving non-relevant ones. [3]

![Figure 1: Information Retrieval System Framework](image)

### 1.3. Information Retrieval Models

**Boolean model:** In the Boolean retrieval model, the indexer module performs a binary indexing in the sense that a term in a document representation is either significant or not. User queries in this model are expressed using a query language that is based on these terms and allows combinations of simple user requirements with the logical operators AND, OR and NOT. Only two possibilities are considered for each document: to be or not to be relevant for the user’s needs, represented by the user query [4] [5].

Vector space model: In this model, a document is viewed as a vector in an n-dimensional document space (where n is the number of distinguishing terms used to describe contents of the documents in a collection) and each term represents one dimension in the document space. A query is also treated in the same way and constructed from the terms and weights provided in the user request. Document retrieval is based on the measurement of the similarity between the query and the documents. Documents with a higher similarity to the query are retrieved by the IRS in a higher position than other documents. [4].

**Probabilistic model:** This model tries to use the probability theory to build the search function and its operation mode. The information used to compose the search function is obtained from the distribution of the index terms throughout the collection of documents or a subset of it. This information is used to set the values of some parameters of the search function, which is composed of a set of weights associated to the index terms [6] [7].

### 1.4. Crawlers

Web crawler is a computer program that browses the World Wide Web in a methodological, automated manner or in an orderly fashion. It visits websites and read their pages and other information in order to create entries for search engine index. It starts with a list of URLs to visit and as crawler visit these URLs it identifies all the hyperlinks in the page and adds them to list of URLs to visit. [12]

A focused crawler or topical crawler is a web crawler that attempts to download only web pages that are relevant to a pre-defined topic or set of topics and avoids the downloading of irrelevant web documents and thus reduce network traffic. [11].

### 1.5. Page Relevancy

Page relevancy of a document retrieved after the user enters his query can be calculated on the basis of the common terms occurring both in document and the query. If the query and document do not have any terms in common, then the document is not considered as relevant. Thus relevancy can also be calculated by the similarity between the two documents by using similarity measures.

### 1.6. Similarity Measures

A similarity measure is a function which computes the degree of similarity between a pair of text objects. Similarity measures rely heavily on terms occurring in both the query and the document. If the query and document do not have any terms in common, then similarity score is very low, regardless of how topically related they actually are. The measures reflect the degree of separation or closeness of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. [8]

#### Types of Similarity Measures

1. **Cosine Similarity**

   When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Given two documents X, Y and their cosine similarity is

   \[
   \cos(x, y) = \frac{X \cdot Y}{|X| \cdot |Y|}
   \]

   Where X, Y m-dimensional vectors are over the term set \(T = \{t1 \ldots tm\}\). Each dimension represents a term with
its weight in the document, which is non-negative. As a result, the cosine similarity is non-negative and bounded between $[0, 1]$. [8]

2. **Jaccard Coefficient**

The Jaccard coefficient, measures similarity as the intersection divided by the union of the objects. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms. The formal definition is:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where $A \& B$ represents the documents and the value of coefficient ranges between 0 and 1. It is 1 when $A = B$ and 0 when $A$ and $B$ are disjoint, where 1 means the two objects are the same and 0 means they are completely different. [8]

3. **Dice Coefficient**

$$s = \frac{2|X \cap Y|}{|X| + |Y|}$$

It is a similarity measure over sets. For sets $X$ and $Y$ of keywords used in information retrieval, the coefficient may be defined as twice the shared information (intersection) over the sum of cardinalities.

4. **Horng and Yeh Coefficient**

In this measure function, the documents are simply ranked. Let $d_1, d_2... d_D$ denote the sorted documents by decreasing order of the values of the similarity measure function, where $D$ represents the number of training documents. The function $r(d)$ gives the relevance of a document $d$. It returns 1 if $d$ is relevant, and 0 otherwise.

The non-interpolated average precision is defined as follows:

$$\text{AvgP} = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{M} \frac{1}{r(d_i)}$$

$r(d_i)$ returns 1, if $d_i$ is relevant and 0 otherwise. [9]

2. **CONCLUSION**

In this paper, it was described that when the user enters the query, page relevancy can be calculated with the use of various similarity measures which is useful in information retrieval by ranking the web pages and providing the more relevant pages earlier than less relevant pages.

**REFERENCES**


