Probability Based Model for Information Retrieval System

Vijay Rana¹, Vijay Dhir²

¹Research Scholar, I.K Gujral Punjab Technical University, Jalandhar, Punjab, INDIA
²St. Solider Institute of Engineering & Technology, Jalandhar, Punjab, INDIA

ABSTRACT

Information retrieval system is designed with the objective to provide the essential utilities, tools, and techniques that assist a user to perform various tasks in finding useful information and knowledge. However, the present indexing techniques have not improved retrieval results, because they only contain the inference information with lack of semantics. Therefore the need for research activities in information retrieval system & enhancements by developing a precisions and intended techniques for the support of heterogeneous infrastructure is apparent. To accomplish this task we proposed Semantic Information Retrieval system that gives an innovative direction for scientific research and pushing information technology to toward making the meaning information and exploits some mathematical and probability techniques to automatically extract valuable information from the particular system. Mathematical model of SIR system is based on the standard principle of probability, where documents ranking is based on their quality and calculates the numerical weight of each document on the basis of citation analysis.

Keywords--- Probability, Information Retrieval, WordNet, Precision, Parameter Estimation.

I. INTRODUCTION

With the rapid growth of information sources existing on the web, it has potential necessary to use some automated techniques to find the valuable information on web and summarized their usage behaviors. These significant facts confer the rising demand of intelligent system, which are able to extract desired information and semantic information retrieval (SIR) system is most promising approach in these circumstances. The vision of SIR system proposed an environment where the information obtainable on the web can be significantly interpreted. It aims to make computer determine the meaning of query on the web pages rather than merely presenting them to users. Semantic Information retrieval is the name of the technique whereby a prospective user of information is able to renovate his requirement for information into an actual list of citations to documents in storage containing information valuable to user.

Although Information Retrieval system is a process of searching, extracting and interpreting queried information from a vast database. It gets a query from a user and provides results with a set of both relevant and non-relevant information’s and also assists users in getting the information they require [8]. In this work we used mathematical approach with the objective to determine and summarize about some phenomenon in the real world that plays a vital role in evaluating the efficiency and performance of the retrieval system. Over the last three eras, probabilistic approaches of information retrieval have been studied comprehensively. In general, these approaches can be classified as techniques of measuring the probability of relevance of information’s to user queries [9]. On the basis of these techniques we argue that the present indexing models have not able to improved retrieval results, because the existing centralized platform and usually a large server can’t ensure the scalability, non-redundancy and interpretability of information provided to users, which results out of semantic heterogeneity. To overcome this problem we propose a new probabilistic approach to information retrieval based upon the ideas and methods of parameter and WordNet based techniques that utilized to assist computers decide the right meaning of the ambiguity information’s and find the most appropriate meaning of ambiguity information according to the context in which it appears. It also used non-parametric and combines information indexing and information retrieval into a hybrid model that combines statistics measure, which are used heuristically in many other retrieval system. We have implemented our model and tested it empirically.
This paper has been broadly divided into five sections. Section 2 gives a brief overview of basic notation of information retrieval system. Third section describes the SIR system 4 sections evaluations of experiment results and 5 sections concludes by presenting open research challenges. The next section describes the basic notation of information retrieval system.

II. BASIC NOTATION

An information retrieval system is an interactive medium to retrieve and manipulate information. Several approaches contribute to its precise development in satisfying the user desired information requirement [1]. In various situations, when a specific search request is presented to a retrieval system, the information’s in its execution can be conceptually described into two forms [3]. First one is total number of relevant documents retrieved and second one is total of irrelevant documents retrieved. Let us start with describing an information retrieval system [12]. By this, we describe the quadruple: \( L = (S, Q, M, \varphi) \), Where \( S \) is a set of documents of an information system of cardinality \(|S|=a\), \( Q \) is a group of queries directed towards the information system of the cardinality \(|Q|=m\), \( M \) is a set of descriptors of the cardinality \(|M|=b\), where as \( \varphi \) is a matching in the form: \( \varphi: Q \to 2^S \).

We utilized the matching \( \varphi \) the algorithm for the assignment of documents to specific queries. For a given query \( q \in Q \) we will call the set \( \varphi(q) \leftrightarrow S \) the reply of the information system. The reply documents made to given query \( q \in Q \)in the order of minimizing suitability of that query’s to the search patterns of documents \( s \in S \) will be classified by Ord \( \varphi(q) \).

Let us assume that we define relation \( R \) of the document that is a binary semantics relation in the form: \( R = (s, m, \mu_s(m), s \in S, m \in M) \). \( \mu_s(m) \colon S \times M \to [0,1] \), is a function specifying for each ordered pair \((s, m)\); \( s \in S, m \in M \), the main significance of the descriptor \( t \in M \) in the description of the document \( d \in S \). On the basis of the binary semantics relation \( R \) in the Cartesian product \( S \times M \), written as \( R \to S \times M \), we can also describe the search pattern of document \( s \in S \) is the semantics set \( R_s \) in the set \( M \) of the descriptors, written as \( R_s \leftrightarrow M \), is the form:

\[
R_s = (m, \mu_{R_s}(m) = \mu_p(s, m) | m \in M) \tag{2}
\]

Further in the condition of a large set \( S \) of documents in an information system, a matter of intended significant is to classified such as algorithm for allocating documents to specifies queries that the time taken for the information system to reply to these queries is acceptable. In many condition excellent of the retrieval approach of documents can be classified by the optimal selection of the value of \( \hat{A} \) with regard to the time and the quality of the retrieval process and then by operating on \( \hat{A} \) level document search patterns.

By the \( \hat{A} \) level search pattern of a given document we will understand the \( \hat{A} \) level semantics set \( R_{s(\hat{A})} \) defined as follows:

\[
R_{s(\hat{A})} = (m, \mu_{F_s(\hat{A})}(m) = \mu_r(m) | m \in R_s(\hat{A})) \tag{3}
\]

Where \( R_s(\hat{A}) \) is the \( \hat{A} \) level set:

\[
R_s(\hat{A}) = (m) | \mu_r(m) \geq \hat{A}, t \in M \tag{4}
\]

The upcoming section describes the role of SIR in information retrieval system.

III. PROPOSED WORK: SEMANTIC INFORMATION RETRIEVAL (SIR) SYSTEM

SIR approach processes a set of associated information in order to retrieve their implicit semantics and attaining most relevant answers. In this paper we proposed three main contributions to find the relevant information: 1st step is parameter based learning approach, 2nd step WordNet based probabilistic approach and 3rd step is define probabilistic models to summarize the probability of relevance of documents for a query.

Step 1: Parameter Estimation Approach

The parameter estimation approach can be classified into three form document learning approach, query learning approach and information learning approach (expressions) [7]. In each of the three approaches, we can differentiate in a two forms: semantics phase and a relevance phase (see figure1). In the semantics phase, we have obtained significance feedback information for a certain modules \( Q_S \times D_S \times E_S \) of \( Q \times D \times E \) (where \( E \) describes the set of expressions in the collection) from which we can obtain probabilistic parameters. These parameters can be utilized in the relevance phase for the development of the information’s of documents and queries.

Document learning approach is orthogonal to the query-information learning approach [2]. In document information learning approach used indexing models to collect relevance feedback information for a specific document \( d \) from a set of queries \( Q_S \) with the set of expressions \( E_S \) occurring in these queries. The parameters obtained from this information can be utilized for the similar document and the same set of expressions \( E_S \).

The key limitation of above two approaches is their restricted application range. In order to tackle these hurdles, we can use information learning approach with specific documents, queries and expressions. A relevance information contains values of features of the objects under consideration (queries, documents and expressions). In the semantics phase, parameters relating to these features are derived from the semantics sample \( Q_S \times D_S \times E_S \). For the relevance phase, there are no restrictions concerning the subset \( Q_S \times D_S \times E_S \) of objects to which these parameters can be applied: new queries as well as new documents and new expressions can be considered.

Step 2: WordNet Based Probabilistic Technique
Several times result of terms on the parameter evaluation approach is not adequate to conclude that information should be activated in the disambiguation. For these instances we used WordNet based probabilistic technique. This measure consists is-a hierarchy of WordNet that based on the idea of information content [6]. The basic idea inside this measure is that two words are more significantly related whenever the quantity of information they distribute in common. The amount of common information of two concepts is identified by the information data of their lowest common subsume (LCS) [10]. Thus, the similarity measure of relationship computation is defined as follows:

\[ \text{Sim}(m_1, m_2) = m \in T(m_1, m_2)^{1-\log(m)} \]  

(5)

Where \( T(m_1, m_2) \) is the group of concepts which subsume equally \( m_1 \) and \( m_2 \) and \( \log(p(m)) \) is the possibility of encountering an illustration of meaning \( m \) in a reference corpus.

In order to describes the probability structure of the WordNet \( T \). The following methods can be utilized, \( c \) is the word to be disambiguated (target word); \( m_1, ..., m_R \) are probable meanings for \( c \); \( s_1, ..., s_j \) are contexts of \( c \) in a WordNet \( T \); \( x_1, ..., x_j \) are words utilized as related characteristics for the disambiguation of \( c \).

In this work the contextual characteristics could be some features such as morphological and semantics, etc), or they could be “surrounding” contextual words of the target word. The contextual characteristics arise in a set point by \( c \), in a context window of specific duration. The context window of size \( e \) will refer obtaining into consideration \( e \) content words to the left and \( e \) content words to the right of the target word. The total number of words obtained into consideration for disambiguation will therefore be \( 2e+1 \).

IV. EVALUATION

The insight into the available literature emphasizes that there are huge number of well classified standards in information retrieval system on which information measures and disambiguation frameworks can be evaluated. Moreover, the analysis constraints have been different [11]. After an appropriately examine of the available evaluation constraints, we have considered Performance Measure and Quality Measure as the most suitable parameters for evaluating SIR. It also summarizes some related techniques, in order to start evaluations between our approaches and other existing ones, to precisely determine the possibility and advantages of our proposed approach.

4.1 Performance Measures

In this section we have described the evolutionary result of the SIR by using some golden standard benchmarks with existing techniques. To evaluate the performance of SIR system we explored the standard precision and recall to evaluate mapping results. These standard attempts to measure the amount of relevant and irrelevant information by evaluating the quantity of the obtained information. In our experiment considered to select only the first 6 and 5 answers for evaluation. A total number of 10 queries in the general and computer engineering discipline are selected for the experiment and it are classified into one categories by the level of search complexity; simple one-word and multiple words queries. In this work we considered ambiguity queries; those queries consists more than one meaning or sense. Table 1 highlights list of 10 queries with their measures and figure 2 defined the precision, recall and f-measure of SIR system on the basis of their relevant and non relevant results.
4.2 Quality Measure

We also measured Precision and Recall for retrieval over three different types of documents: text only, text with semantic ambiguity and text with annotation markup that has been augmented by inference. We also ran queries over documents enhanced by means of gathering information from external sources such as WordNet. We evaluated SIR system using the TREC evaluation package available from the TREC Web site.

V. CONCLUSION

This paper emphasizes the technologies contributing most towards the next generation of information retrieval system and also conducts a review of research in the area of probability measure and identifies the major theoretical and practical issues which need to be addressed. The exponential advancements in semantic information retrieval (SIR) system have enabled users to experience enhanced delivery of personalized services & information through the integration of various existing technologies such as WordNet. The probable advantages of SIR system can be applied to extracting knowledge oriented information from the web. The future work of SIR system will be the integration of this technique with a system that deals with the semantic enhancement of ambiguity information and serves its users more intelligently. It gives a new way for logical research, impelling the approaches toward controlling the meaning of information and making an intelligence component that can accurately executes their jobs.

REFERENCES


We used two types of inference to enhance information retrieval: reasoning over ambiguity instances and reasoning over the semantic hierarchy. DAML being a machine understandable language, WordNet measure can support in information development that provides answering queries beyond the possibility of large document hierarchy. As an instance obtaining the results of a Movie from the description of movie showings, a WSD can obtain more information about the movie from the MDB site and detail the film showing result outline. A query searching for movies of the type Action movie can thus be satisfied, therefore the initial event explanation was not sufficient for the task. development and comparative experiments, International Journal of Information Processing and Management, pp779-808.