A Framework for Efficient Information Retrieval Using NLP Techniques

R. Subhashini[1] and V. Jawahar Senthil Kumar[2]

[1] Research Scholar, Sathyabama University, Chennai-119, India
subhaagopi@gmail.com
[2] Assistant Professor, Anna University, Chennai, India
veerajawahar@yahoo.com

Abstract. In the academic area, the Internet is used as a scientific resource. However, finding appropriate information on the Web remains difficult. To simplify this process, we designed the Information retrieval framework to retrieve information from the Web Using NLP Techniques. Many information retrieval systems are based on vector space model (VSM) that represents a document as a vector of index terms. To remedy this problem, documents and queries are optimized using NLP. In this paper, the architecture of the proposed tool is presented and it proposes a new approach over the traditional VSM by considering only the nouns and verbs of the documents extracted from NLP as the constituting terms for the VSM instead of the traditional term "word". Such a mechanism may raise the effectiveness of the Information Retrieval by increasing the evaluation metrics values.

Keywords: NLP, Text Mining, Information Retrieval, Vector Space Model.

1 Introduction

Many Natural Language Processing (NLP) techniques [5], including stemming, part of-speech tagging, compound recognition, de-compounding, chunking, word sense disambiguation and others, have been used in Information Retrieval (IR). The core IR task we are investigating here is document retrieval. Several other IR tasks use very similar techniques, e.g. document clustering [2], filtering, new event detection, and link detection, and they can be combined with NLP in a way similar to document retrieval. Today, the amount of documents published on the internet grows dramatically, especially in education and e-learning activity and if the same information retrieval methods is used, then the time to find relevant document will also increase the same way. One possible solution is to develop software framework, which would improve the quality of current search engines and decrease time dedicated to searching process.

2 System Architecture

Our proposed system sends the optimized query to the search engine API [4]. This search engine API is a web service and it collects the top n documents from the search

V.V. Das, J. Stephen, and Y. Chaba (Eds.): CNC 2011, CCIS 142, pp. 391–393, 2011.

engine like yahoo. The optimized query obtained by NLP is compared with the optimized retrieved documents by cosine similarity measure [3] and it finds relevant documents to that query and displays them as a ranked list according to their similarity score. The Fig. 1 shows the architecture of the proposed system. The uniqueness of our proposed method to the existing method is that it calculates the similarity measure for only the nouns and verbs of the query and documents optimized by NLP by using sharp NLP [1]. We used Sharp NLP parser to parse the English sentences in to subject, verb and object. The noun phrases and verb phrases are extracted from the parser and they are used for calculating the similarity measures.

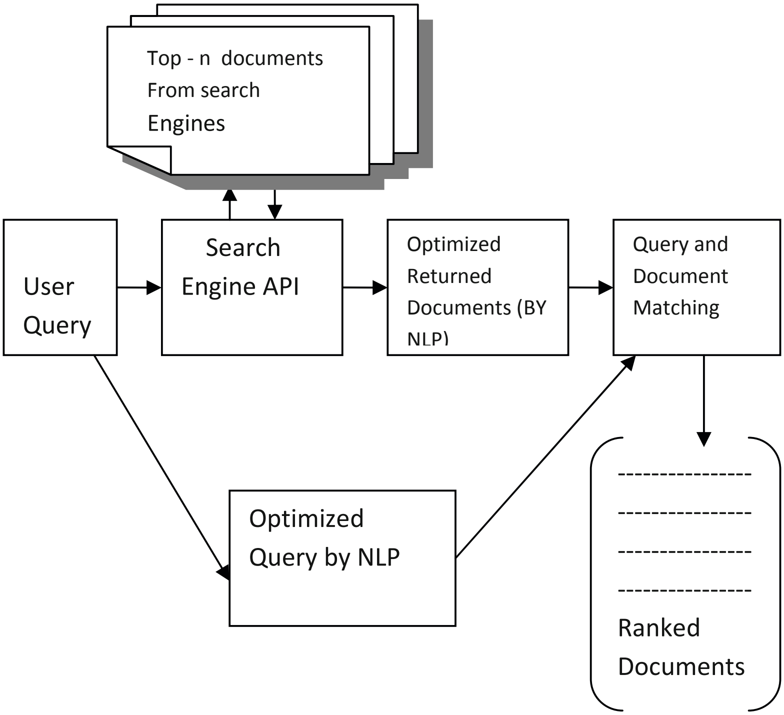

Fig. 1. Architecture of the proposed System

3 Experimental Setup and Evaluation Measures

Traditional methods used text datasets but in this system, web documents are collected using yahoo API and similarity measures are implemented on it. BOSS (Build your Own Search Service) is Yahoo open search web services platform. The application was written in c#. The query is given to yahoo API and it returns the top n documents i.e., the xml documents and it contains the URL's. Here, the value of n is set to 10 i.e., it represents the first 10 documents retrieved by the search engine. With that URL's the web pages are retrieved, optimized and then compared with the optimized query to get the relevant output. The relevant documents were ranked based on their similarity score. The system is tested with 50 queries which is represented as 5 sets