58

11/21

NOVEL APPROACH TO SEARCH DOCUMENTS USING LLH DISTANCE CALCULATION

[1]Mrs.T.Yogameera, [2]Dr.D.Shanthi

[1]Assistant Professor, Department of Computer Science and Engineering, Nadar Saraswathi College of Engineering and Technology, Theni, Tamil Nadu.

[2] Professor and Head, Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu.

E-mail: [1] yogmeeraarasu@gmail.com, [2]dshan71@gmail.com

ABSTRACT- Data management is one of the hectic tasks that the server end faces. Surplus amount of documents are pooled in the cloud from where the target hyper documents is expected to be retrieved back within the fraction of time when requested for, hence it is mandatory to organize and store the data such that it is easily searchable and reachable across domains. In this view Natural Language Processing is the necessary step to ease the web user task; query normalization is the primary requirement for optimizing the search. It includes multiple phases like Language Identification, Tokenization, Stemming, Lemmatization, Sentence Breaking, parts of Speech identification, Tagging, Chunking and Syntax Parsing, However this paper concentrates on the lemmatization techniques and edit distance calculation in the query-text normalization phase. The root or base form (lemma) of a query token is identified following certain rules of the Levenshtein distance threshold in the pre-phase. The optimized keyword or lemma for the searched documents identified. The ranking of matched results using the hamming distance calculation is performed to get an increased precise result. Hence the proposed LLH (Levenshtein-Lemmatization-Hamming) method is the next step towards Natural Language processing.

KEYWORDS—Lemmatization, Hamming distance, Levenshtein distance, Query optimization, Natural language processing.

1. INTRODUCTION: When a request for the document is posted on the web by the user, it is by default that queries need to be preprocessed before entering into the search phase. Query-Word normalization is performed aiming to identify the user need at core. Stemming and lemmatization are the two popular approaches used after tokenization step to generate proper keywords.

Edit distance calculation plays an important role in text comparison. The justifiable change of query word to match the index paves ways for a better indexed search. Sequentially matching all the index tags is a resource and time consuming process. We have used two distances calculation steps one at the pre-phase to start the index search and at the end to rank the documents.

Stemming is a process where the base word is identified by just pruning the inflectional endings of the terms like prefixes and suffixes. The Poster Stemming algorithm and Lancaster stemming, stop word removal, suffix stripping and other non English stemmers are used in this process. For example, when the word "SEARCHABLE" is given the suffix part "able" is pruned, resulting in stem "SEARCH". However there is no semantic analysis for the words in stemming. It just follows the stripping rules to

Mani