Certain Investigations on Improving Ranking and Grouping Mechanisms of Search Results using Web Mining Techniques

Abstract

Web mining is broadly classified into three categories namely, Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). WCM analyses the content of web pages using different mining techniques. WSM finds the popularity score of a page by following link structure of the web. WUM analyses the data about the web users to predict future usage of the web. The thesis aims at improving the ordering and grouping of search results using WSM and WCM so that the user can select the relevant document in a faster manner.

The first part of the thesis aims at improving the existing WSM based PageRank algorithm by including weight for the links between the connected documents. For weight calculation, distance among the documents is calculated using Euclidean distance measure. Using the calculated distance values, weight values are computed. These weight values represent the similarity among the linked documents. In addition to the computation of weight values using Euclidean distance measure, semantic similarity values among the documents are calculated using Wu and Palmer measure and included in the existing PageRank formula.

In the first part of the work, initially distance values are computed and from the distance values, similarity values are computed and these similarity values are used as the weight for the links. Instead of computing weight values from the

distances, second part of the thesis suggests a method of calculating weight using cosine similarity measure. Also semantic similarity has been combined with cosine similarity values.

Even after calculating the popularity score values using different techniques, it is very difficult for the user to select a relevant web page if it is present in the middle of the returned search results. This problem has been addressed in the other parts of the thesis by clustering the web pages and displaying the clusters in a single page to make it easy for the user to select the relevant page.

For clustering web pages, contents from specified tags are extracted and optimized K-means clustering has been used. Optimization of K-means is done by selecting the cluster centroids using scale factor method and the number of clusters by knee finding method. Since this method took longer time for computing the centroids and the number of clusters, next work improves the existing Suffix Tree Clustering (STC) algorithm by computing semantic similarity relationship between the terms in different documents.

The proposed suffix tree based clustering method restructures the documents by computing the semantic similarity value among the terms in different documents. Once the documents are restructured, suffix tree is constructed. From the suffix tree, base clusters are identified and by removing common base clusters and the clusters with all the documents final clusters are formed. The problem occurred in this method is the selection of number of clusters. This problem has been avoided by the next method using existing Lingo algorithm along with semantic similarity values.

The proposed lingo algorithm based clustering method induces the cluster labels and the documents are assigned to the respective clusters. Cluster labels are induced by finding frequent phrases from the collection of documents. Documents are assigned to the induced cluster labels by computing cosine similarity between the documents and the cluster labels. In this method, number of clusters is selected randomly. Instead it is computed from the term-document matrix.

The performance of the proposed popularity score calculation algorithms has been done using the measures precision, recall and f-measure. Experimental results indicate that effectiveness of the proposed methods over the existing link analysis methods.

The performance of the proposed clustering algorithms has been performed using intra-cluster distance and inter-cluster distance. The results show the effectiveness of the proposed methods over the existing clustering algorithms.