Mining Unstructured Data from Web Using Soft Computing

Ms. Sonali A. Deshpande, Dr. P.R. Deshmukh
Sipna’s COET, Amravati, HOD, Sipna’s COET, Amravati
sonali21.deshpande@gmail.com

Abstract

Copious material is available from the World Wide Web (WWW) in response to any user-provided query. It becomes tedious for the user to manually extract real required information from this material. Large document collections, such as those delivered by Internet search engines, are difficult and time-consuming for users to read and analyse. The detection of common and distinctive topics within a document set, together with the generation of multi-document summaries, can greatly ease the burden of information management. In the paper, a technique is proposed called Mining Unstructured Data From Web Using Soft Computing that creates the clusters of web documents using fuzzy clustering which focuses on this problem of mining the useful information from the collected web documents.

Keywords--Search Engine, Web documents, Fuzzy Clustering, Soft computing

1. Introduction:

WWW is a huge repository of information consisting of hyperlinked documents spread over the internet. For a user, it is practically impossible to search through this extremely large database for the information needed by him. The search engine uses crawlers to gather information and stores it in database maintained at search engine side. For a given user’s query the search engine searches in the local database and very quickly displays the results. The huge amount of information is retrieved using data mining tools. Classification, Clustering and Association tools etc. are used for data mining technique. Clustering plays a key role in searching for structures in data. As the number of available documents nowadays is large, hierarchical approaches are better suited because they permit categories to be defined at different parness levels. The problem of clustering in finite set of data is to find several cluster centers that can properly characterize relevant classes of finite set of data such that degree of association is strong for data within blocks of the partition and weak for data in different blocks. When the weakness of a crisp partition of finite set of data is replaced with a fuzzy partition, this area is known as fuzzy clustering. Fuzzy clustering is a relevant technique for information retrieval. As a document might be relevant to multiple queries, this document should be given in the corresponding response sets, otherwise, the users would not be aware of it. Fuzzy clustering seems a natural technique for document categorization. There are two basic methods of fuzzy clustering, one which is based on fuzzy c-partitions, is called a fuzzy c-means clustering method and the other, based on the fuzzy equivalence relations, is called a fuzzy equivalence clustering method. The purpose of this paper is to propose a search methodology that consists of how to find relevant information from WWW. In this paper, a method is being proposed of document clustering, which is based on fuzzy equivalence relation that helps information retrieval in the terms of time and relevant information.

2. Literature Review:

Document clustering has been widely applied in the field of information retrieval for improving search and retrieval efficiency [1]. Furthermore, document clustering has also been applied as a tool for browsing large document collections [2] and as a post-retrieval tool for organizing Web search results into meaningful groups [3]. Document clustering is recently applied to dynamically discover content relationships in e-Learning material based on document metadata descriptions [4, 5]. Main focus is on the discovery and representation of unobvious or unfamiliar knowledge about a domain rather than on facilitating the access to specific information resources through a set of document clusters. Various techniques for accurate clustering have been proposed [6], e.g. K-MEAN [7, 8], CURE [9], BIRCH [10], ROCK [11]. K-MEAN clustering algorithm is used to partition objects into clusters while minimizing sum of distance between objects and their nearest center. In statistics and machine learning, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. CURE (Clustering Using Representation) represents clusters by using multiple well scattered points called representatives. A constant number ‘c’ of well scattered points can be chosen from 2c’ scattered points for merging two clusters. CURE can detect clusters with non-spherical shapes and
works well with outliers. CURE achieves this by representing each cluster by a certain fixed number of points that are generated by selecting well scattered points from the cluster and then shrinking them toward the center of the cluster by a specified fraction. Having more than one representative point per cluster allows CURE to adjust well to the geometry of non-spherical shapes and the shrinking helps to dampen the effects of outliers. To handle large databases, CURE employs a combination of random sampling and partitioning. A random sample drawn from the data set is first partitioned and each partition is partially clustered. The partial clusters are then clustered in a second pass to yield the desired clusters.

BIRCH (Balance and Iterative Reducing and Clustering Hierarchies) is useful algorithm for data represented in vector space. It also works well with outliers like CURE [12]. BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering with the available resources (i.e., available memory and time constraints). BIRCH can typically find a good clustering with a single scan of the data, and improve the quality further with a few additional scans. BIRCH is also the first clustering algorithm proposed in the database area to handle “noise” (data points that are not part of the underlying pattern) effectively.

However, the traditional clustering algorithms fail while dealing with categorical attributes. As they are based on distance measure so their merging processing is not accurate in case of categorical data.

ROCK (Robust Clustering Algorithm for Categorical Attributes) gives better quality clusters involving categorical data as compared with other traditional algorithms.

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster.

In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

### 3. Analysis of Problem:

The growth of the World Wide Web has enticed many researchers to attempt to devise various methodologies for organizing such a huge information source. Scalability issues come into play as well as the quality of automatic organization and categorization. Documents on the web have a very large variety of topics, they are differently structured, and most of them are not well-structured. The nature of the sites on the web varies from very simple personal home pages to huge corporate web sites, all contributing to the vast information repository. The real motivation behind this work is to help in the resolution of this problem by taking one step further toward a satisfactory solution. The intention is to create a system that is able to categorize web documents effectively, based on a more informative representation of the document data, and targeted towards achieving high degree of clustering quality by implementing fuzzy clustering algorithm. So the motivation behind this work is as follows:

- The major source of information is from the Web.
- Traditional information retrieval (IR) approaches are hardly appropriate due to enormous size.
- Manually extracting real required information is difficult to read and analyze.
- Many search engines gives a long list of ranked documents and most of them are irrelevant.
- Typical queries retrieve hundreds of documents; most of them have no relation with what the user is looking for.
- So, the main problem is regarding organization of document data.
- One of the Solutions to this is clustering

### 4. Proposed Work & Objectives:

The work focuses on mining of the useful text as per the user query from the collected web documents. The dissertation work will be carried out as follows.

1. The application domain will be selected.
2. Different web documents will be collected from the World Wide Web related to the selected application domain.
3. Documents will be preprocessed for removing frequently used stop words.
4. Rules will be generated as per the association in the words of query provided by user. These rules will be added into rule base.
5. Number of clusters will be created using fuzzy clustering algorithm.
6. The clusters will be selected as per user query for mining unstructured data.
7. The text will be mined from documents as per selected rules and clusters.
5. Desired Implications:

Search engines were introduced to help find the relevant information on the web, such as Google, Yahoo!, and AltaVista. However, search engines do not organize documents automatically; they just retrieve related documents to a certain query issued by the user. While search engines are well recognized by the Information Retrieval community, they do not solve the problem of automatically organizing the documents they retrieve. The problem of categorizing a large source of information into groups of similar topics is still unsolved. With the proposed concept this problem can be overcome at some level this can be implemented as Query provided by user will be fired on the collected web pages from World Wide Web (ie. On already formed clusters of the web pages), related to the application domain to retrieve (mine) the text/data. The text (data) will be mined from documents (or web pages) as per the selected rules and already formed clusters.

6. References: