Lingo an approach for Clustering

Poonam C. Fafat
Department of Information Technology
PRMIT&R,Badnera

Prof.S.S.Sikchi
Department of Information Technology
PRMIT&R,Badnera

Abstract— **Clustering** is a process of forming groups (clusters) of similar objects from a given set of inputs. When applied to web search results, clustering can be perceived as a way of organizing the results into a number of easily browsable thematic groups. Top Down clustering is a type of Hierarchical Clustering. It tries to find bigger clusters first and then does fine grained clustering on these clusters. Hence the name Top Down. Any clustering algorithm can be used to perform the Top Level Clustering (finding bigger clusters) and the Bottom Level Clustering (fine grained clustering on each of the top level clusters).

Keywords- Top down Clustering;snippets;outlier detection

I. INTRODUCTION

Top Down clustering is a type of Hierarchical Clustering. It tries to find bigger clusters first and then does fine grained clustering on these clusters. Hence the name Top Down. Any clustering algorithm can be used to perform the Top Level Clustering and the Bottom Level Clustering The first step to execute the top down clustering, would be to run clustering algorithm , preferably with clustering parameters which will produce bigger clusters. This would be the top level clustering. Then, the output of this clustering should be post processed, to group them into respective top level clusters. **Clustering** is a process of forming groups (clusters) of similar objects from a given set of inputs. When applied to web search results, clustering can be perceived as a way of organizing the results into a number of easily browsable thematic groups. Moreover, the advanced users, who sometimes issue general queries to learn about the whole spectrum of sub-topics available, will no longer be made to manually scan hundreds of the resulting documents. Instead, they will be presented with a concise Summary of various subjects, with an option of drilling down selected topics.What should be emphasized about web search clustering is that the thematic groups must be created *ad hoc* and fully automatically. Additionally, in contrast to the classic Information Retrieval tasks, the web search clustering algorithms do not operate on the full text documents, but only on the short summaries of them returned by search engines, called **snippets**. Therefore, the algorithms must take into account the limited length and often low quality of input data.

II. BACKGROUND

Compared to classic full-text clustering methods, search results clustering has its own characteristics: (1) its processing targets are document "snippets" returned by search engine rather than full-text sources. Therefore, the algorithms must be prepared for limited length and often low quality of input data. (2) ad-hoc processing. This type of clustering should be regarded as the post-processing part of the on-line search engine. Therefore, the algorithms should satisfy the need of real-time web service. (3) the goal of search results clustering is to help the users to locate the interest more quickly. Therefore, the algorithms should be capable of supplying users with meaningful, concise and unambiguous clustering labels.

In this paper the clustering of data will be done using the lingo algorithm. Lingo algorithm is a clustering algorithm.

III. PROPOSED WORK

The proposed work is based on the clustering algorithm. To the desired search engine the input query will be given.

Figure 1.1

1) Implementation of Clustering Algorithm on dataset which is high dimensional categorical data.

2) Implementation of Detection of Outliers high dimensional categorical data.

3) Performance analysis of Lingo algorithm with other clustering method.

Outlier detection is a fundamental task that is useful in a number of data analytic applications. Outliers arise due to mechanical faults, changes in system behaviour, fraudulent behaviour, network intrusions or human errors. It deals with the problem of identifying rare or a typical points that widely diverge from the general behaviour or model of the

data. The process of detecting outliers and subsequently using them for data analysis relies on the underlying application. For example, outlier detection can be employed as a pre-processing step to clean the data set from erroneous measurements and noisy data points. On the other hand, it can also be used to isolate suspicious or interesting patterns in the data. Examples include fraud detection, customer relationship management, network intrusion, clinical diagnosis and biological data analysis.

Clustering is a process of forming groups (clusters) of similar objects from a given set of inputs. Good clusters have the characteristic that objects belonging to the same cluster are "similar" to each other, while objects from two different clusters are "dissimilar". The idea of clustering originates from statistics where it was applied to numerical data. However, computer science and data mining in particular, extended the notion to other types of data such as text or multimedia.

Clearly, web snippets belong to the class of textual data and hence it is the text clustering algorithms that should be regarded as a base for the web search results clustering systems.

When designing a web clustering algorithm, special attention must be paid to ensuring that both contents and description (labels) of the resulting groups are meaningful to the users. The majority of currently used text clustering algorithms follow a scheme where cluster content discovery is performed first, and then, based on the content, the labels are determined. Unfortunately, this may result in some groups' descriptions being meaningless to the users, which in turn, is very often caused by the nonsensical content of the clusters themselves. To avoid such problems LINGO adopts a radically different approach to finding and describing groups.

The general idea behind LINGO is to *first* find meaningful descriptions of clusters, and then, based on the descriptions, determine their content. To assign documents to the already labeled groups LINGO could use the Latent Semantic Indexing in the setting for which it was originally designed: given a query – retrieve the best matching documents. When a cluster label is fed into the LSI as a query, as a result contents of the cluster will be returned. This approach should take advantage of the LSI's ability to capture high-order semantic dependencies in the input collection. In this way not only would documents that contain the cluster label be retrieved, but also the documents in which the same concept is expressed without using the exact phrase. In web search results clustering, however, the effect of semantic retrieval is sharply diminished by the small size of the input web snippets This, in turn, severely affects the precision of cluster content assignment. That is why we have decided to use the simple Vector Space Model instead of the LSI to determine the cluster content.

To become a full-featured clustering algorithm, the process of finding cluster labels and contents must be preceded by some preprocessing of the input collection. This stage should encompass text filtering, document's language recognition, stemming and stop words identification. It is also recommended that post-processing of the resulting clusters be performed to eliminate groups with identical contents and to merge the overlapping ones. As a summary of the introductory discussion, in Figure 1.2 we present the main phases of LINGO.

```
/** Phase 1: Preprocessing */
        for each document
{
            do text filtering;
          identify the document's language;
            apply stemming;
            mark stop words;
  }
  /** Phase 2: Feature extraction */
 discover frequent terms and phrases;
 /** Phase 3: Cluster label induction */
 use LSI to discover abstract concepts;
        for each abstract concept{
        find best-matching phrase;
}
      prune similar cluster labels;
/** Phase 4: Cluster content discovery */
          for each cluster label{
use VSM to determine the cluster contents;
}
  /** Phase 5: Final cluster formation */
          calculate cluster scores;
           apply cluster merging
```

Noteworthy is the fact that all five phases of LINGO are independent and easily separable. This allows to manipulate the quality and resource requirements of the algorithm by providing alternative implementations of some of its components.

Conclusion

The aim of this project is to experiment with an effective clustering algorithm on the search results returned from "Google" search engine. The clustering method used in this project is called "Description Comes First (DCF)". The core idea is to first find meaningful cluster labels, then assign snippets to them to create proper clusters. So the algorithm can be split into two major phases: cluster label induction phase and cluster content assignment phase.
Evaluation of the algorithm will be carried out on the test data from Open Directory Project (ODP) which serves as a source of pre-clustered document snippets.

The main result of this project is the design of LINGO – a description-oriented web search results clustering algorithm. We strongly felt that it should be easier to find documents that match a meaningful cluster label than to describe a potentially senseless group of snippets. To verify the practical value of our idea we implemented LINGO as a component of the Carrot2 framework. During the implementation work, we identified additional factors that significantly influence the quality of clustering. We claim that proper preprocessing of the input data, including language identification, is of crucial importance in web mining tasks.

References

[1] P. Andritsos, P. Tsaparas, R. Miller, and K. Sevcik, "LIMBO: Scalable Clustering of Categorical Data," Proc. Ninth Int'l Conf. Extending Database Technology (EDBT '04), pp. 123-146, 2004.

[2] R. Ng and J. Han, "CLARANS: A Method for Clustering Objects for Spatial Data Mining," IEEE Trans. Knowledge and Data Eng., vol. 14, no. 5, pp. 1003-1016, Sept./Oct. 2002.

[3] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," Information Systems, vol. 25, no. 5, pp. 345-366, 2001

[4] J. Basak and R. Krishnapuram, "Interpretable Hierarchical Clustering by Constructing an Unsupervised Decision Tree," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 1, Jan. 2005

[5] Eugenio Cesario, Giuseppe Manco, Riccardo Ortale, "Top-Down Parameter-Free Clustering of High-Dimensional Categorical Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 12, DECEMBER 2007.

[6] D. Barbara´, J. Couto, and Y. Li, "COOLCAT: An Entropy-Based Algorithm for Categorical Clustering," Proc. 11th ACM Conf. Information and Knowledge Management (CIKM '02), pp. 582-589, 2002.

.