Online Clustering of Linguistic Data

Shallu Bajaj[1] MS Sonia[2]
[1]A.P in CSE and IT Department[2]PGScholar
[1, 2]Kurukshetra Institute of Technology and Management,
[1, 2]Kurukshetra University, Kurukshetra

Abstract—*Cluster linguistic data requires addressing the issues on linguistic similarity. The specific instance of linguistic data we want to focus on is news stories. In study of news stories we find that News stories are having their on linguistic notation generally used and while in clustering its require to change the clusters over the time as new news comes. Due to the huge size of news datasets, it is not feasible to store all of the data. We only focus to find a appropriate approach for clustering. In our algorithm, we are using vector in high dimensional vector space model (VSM) for representing documents and clusters and then used cosine distances between vectors. Linguistic data is ubiquitous.*

Keywords: -Linguistic data, news story clustering, Vector Space Model (VSM) , ubiquitous

I. Introduction

Data Mining is described as a process of discover or extracting interesting knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouses…etc. This knowledge contributes a lot of benefits to business strategies, scientific, medical research, governments and individual. Data mining, *the extraction of hidden descriptive or predictive information from large databases*, is a powerful new technology with great potential to help companies and data analysts focus on the most important information in their data repositories. This is also sometimes referred to as Knowledge Discovery from Data (KDD). So KDD is: 'the non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data'.

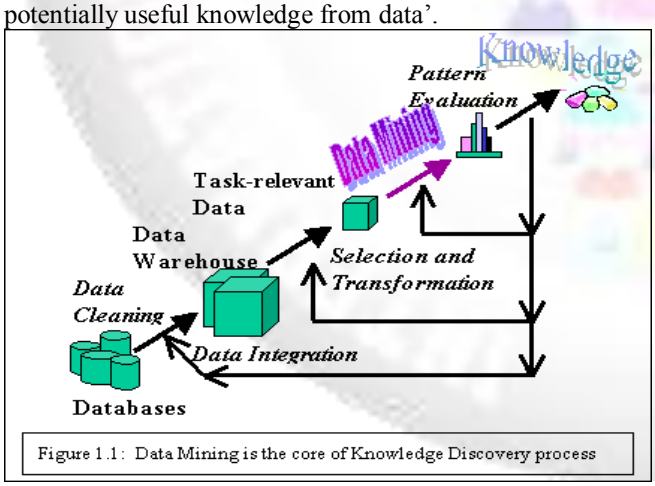

Fig. 1: Architecture Data Mining

A wide range of companies has deployed successful applications of data mining. While early adopters of this technology have tended to be in information-intensive industries such as financial services and direct mail marketing, the technology is applicable to any company looking to leverage a large data warehouse to better manage their customer relationships. Two critical factors for success with data mining are: a large, well-integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied (such as customer prospecting, retention, campaign management, and so on).

II. Linguistic Data Clustering

Data mining discovers description through clustering, visualization, association, sequential analysis. Clustering is a primary data description method in data mining which groups most similar data. Data clustering is a common technique for data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait. Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects, which are "similar" between them and are "dissimilar" to the objects belonging to other clusters fig. 2 shows the process of clustering with a simple graphical example:

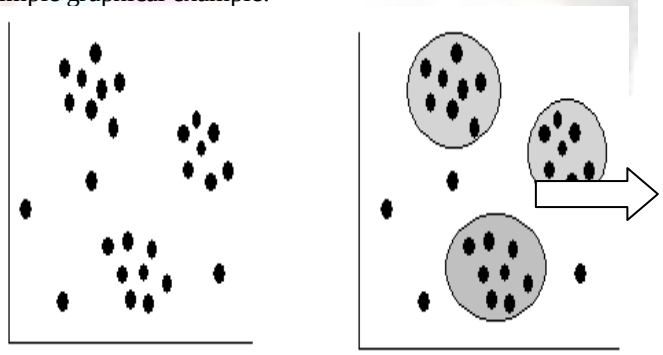

Fig 2: Clustering of data

. Depending on the clustering technique, clusters can be expressed in different ways:

- Identified clusters may be exclusive, so that any object belongs to only one cluster.
- They may be overlapping; an object may belong to several clusters.
- They may be probabilistic, whereby an object belongs to each cluster with a certain probability.

Clusters might have hierarchical structure, having crude division of objects at highest level of hierarchy, which is then refined to sub-clusters at lower levels

A. Clustering Approaches

There exist a large number of clustering algorithms in the literature. The choice of clustering algorithm depends both on the type of data available and on the particular purpose

and application. In general, major clustering methods can be classified into the following categories:-

B. Partitioning algorithms:

K-Means algorithm is one of the partitioning based clustering algorithms. K-Means clustering is an algorithm to cluster or to group data objects based on attributes/ features into k number of groups where k is positive integer number and should be given initially. The grouping is done by minimizing the sum of squares of distances between data and the cluster centroid where centroid is mean value of the cluster.

Let $X=\{x_i \mid i=1, 2, \ldots\ldots n\}$ be a data set with n objects, k is the number of clusters, m_j is the centroid of cluster cj where j=1, 2,…k. Then the algorithm finds the distance between a data object and a centroid by using the following Euclidean distance formula.

Euclidean distance formula= $\sqrt{| x_i - m_j|^2}$

Starting from an initial distribution of cluster centers in data space, each object is assigned to the cluster with closest center, after which each center itself is updated as the center of mass of all objects belonging to that particular cluster.

C. Hierarchy algorithms:

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups close to one another, until all of the groups are merged into one. The divisive approach, also called the top-down approach, starts with all the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster.

III. Representation Model

A. Vector space model :

The most popular representation model used in information retrieval and text mining .In VSM, a text document is represented as a vector of terms.

<t1, t2, ...,ti, ..., tn>.

Each term ***ti*** represents a word or a phrase. A set of documents is represented as a set of vectors that can be written as a matrix. Where each row represents a document, each column indicates a term, and each element *xji* represents the frequency of the ith term in the Jth document.

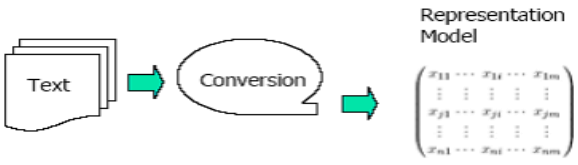

Fig. 3:

In this, each set of document terms was filtered from stop-words and had their suffix removed using the stemming routine. The terms were then represented in vector space using TF.IDF weighting. Document-cluster similarities were found by calculating the cosine similarity between the document and the centroid vector of the relevant cluster. To filter insignificant words, it is possible to use a list of stop words. The list contains common and function words like the, you, an etc. Stop words are language specific, but sometimes a list may even contain words to work better in specific topics. Using a list of stop words can reduce the complexity and improve performance.

B. Word stemming:

Word stemming reduces word to their so called root or base. The advantage of doing this in information retrieval is to reduce the data complexity and size. Word stemming is often used as query broadening in search systems. Porter's suffix stemming routine was released in 1980, and it became the standard algorithm for suffix stemming in the English language. Suffix stemming of words is the procedure to stem words to their root, base or stem form. Example given the words**:**

HAPPEN
HAPPENS
HAPPENED
HAPPENING

A stemmer automatically removes the various suffixes –ED, -ING, –S to leave the base term happen.

C. Term frequency-inverse document frequency:

Term frequency–inverse document frequency (tf–idf) is one of the most commonly used term weighting schemes in information retrieval systems. Tf-idf is used to determine the importance of terms for a document relative to the document collection it belongs to. The tf-idf product is calculated with the term frequency (tf) and the inverse document frequency. The term frequency function tf (t, d) where t is the term and the d is the document. The tf can be represented as the raw frequency f(t, d), that is the number of times t is in document d. After filtering the high frequency words it reconstructs the document, which now consists of very fewer words than the original document. Now this document needs to be converted to some vector (using Vector Space Representation) form for the purpose of calculate some distance with other document. Each possible word would represent a dimension in this vector space. If an article contained a word, it would have 1 added to that dimension. At the end, the vector would be normalized.

IV. Conclusion

In this paper we have presented the news data .To filter insignificant words; it is possible to use a list of stop words. The list contains common and function words like the, you, an etc. Stop words are language specific, but sometimes a list may even contain words to work better in specific topics. Using a list of stop words can reduce the complexity and improve performance. The advantage of doing this in information retrieval is to reduce the data complexity and size.

REFERENCES

[1] Lev Reyzin,"Online Clustering of Linguistic Data," Princeton University Class of 2005, BSE Junior Independent Work Advised by Professor Moses Charikar.
[2] Martina Naughton, Nicholas Kushmerick, and Joe Carthy,"Clustering sentences for discovering events in news articles," School of Computer Science and Informatics, University College Dublin, Ireland *f*martina.naughton, nick, joe.carthy*g*@ucd.ie.
[3] Grigore, M., "Introduction to Stemming," 2008.
[4] Peter D. Turney, Patrick Pantel ,"From Frequency to Meaning: Vector Space Models of Semantics," Journal of Artificial Intelligence Research 37 (2010) 141-188
[5] TayebKenaza, Abdelhalim Zaidi "Clustering approach for false alerts reducing in behavioral based intrusion detection systems**"** Volume 4 Issue 5 IEEE 2010.
[6] Chang-Tien Lu, Arnold P. Boedihardjo, PrajwalManalwar "Exploiting Efficient Data Mining Techniques to Enhance Intrusion Detection Systems**"** Volume 2 Issue 3 IEEE 2005
[7] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir "Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification**"**2011 7th International Conference on IT in Asia (CITA).
[8] Cuixiao Zhang; Guobing Zhang; Shanshan Sun "A Mixed Unsupervised Clustering-based Intrusion Detection Model**"** 2009 Third International Conference on Genetic and Evolutionary Computing.