Research Paper

Document clustering for Information Retrieval – A General Perspective

P.Prabhu
Assistant Professor in Information Technology, DDE
Alagappa University, Karaikudi,
Tamilnadu, India

Abstract

Information Retrieval (IR) is an emerging subfield of information science concerning representation, storage; access and retrieval of information .Current research areas within the field of IR include searching and querying, ranking of search results, navigating and browsing information, optimizing information representation, storage, document classification and clustering. The primary objective of this paper is to understand the method of using document clustering to improve their information retrieval. This paper first discussed method for clustering documents for information retrieval in easy steps by introducing various types of web/electronic repositories. Second explains the steps involved for preprocessing the documents. Second Clustering method especially k-means algorithm is discussed for clustering documents.

Keywords: Information Retrieval, document clustering, document classification.

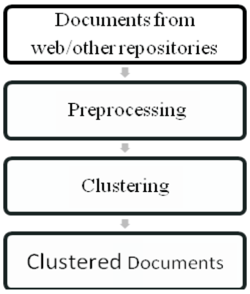
I Introduction

The study of information retrieval is not new to computer science. Traditionally, information retrieval was a manual process; mostly happening in the form of book lists in libraries, and in the books themselves, as tables of contents, other indices etc. These lists/tables usually contained a small number of index terms (e.g. title, author and perhaps a few subject headings) due to the tedious work of manually building and maintaining these indices. Today, information retrieval plays a much larger part of our everyday lives. In modern information retrieval systems, several models exist to represent the information contained in a large collection of textual documents. Within information retrieval, clustering (of documents) has several promising applications, all concerned with improving efficiency and effectiveness of the retrieval process. Most IR systems are based on inverted indices, which, for each keyword in the language, store a list of documents containing that keyword. The Boolean model for information retrieval is a simple retrieval model based on set theory and Boolean algebra. In its essence, the Boolean representation of a document is a set of terms, where the terms are words from the document extracted using different measures such as filtering. The Vector Space model provides a different way of looking at the same information, but is not used as often in practice. A vector space implementation stores a lists of (keyword, frequency) pairs for each document in the data set. This allows a set of documents to be visualized as points in an n dimensional space, where n is the total number of keywords in the language. The applications of Document clustering in information Retrieval include finding similar documents, search result clustering and faster and better searching.

II Document Clustering

It is a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval. Clustering documents involves four stages. Each stage has multiple sub stages. Today, information retrieval plays a much larger part of our everyday lives – especially with the advent of the Internet, and the World Wide Web (the Web) in particular. During the last 10 years, the amount of information available in electronic form on the Web has grown exponentially. Almost any kind of desired information is available on the Web, including: Bibliographic collections, news and message files, software libraries, multimedia repositories, online encyclopedias, commercial information etc. Cluster documents to allow users to better preview and navigate the information structure of the returned results and organise documents into a predefined hierarchy of categories, where the user benefits from

familiar terms when navigating to the right information. The following Figure 1 describes the steps involved in document clustering.



Clustering methods depend on various preprocessing techniques to achieve optimal quality and performance. Selecting the best preprocessing methods for a given clustering algorithm is almost an art, but we will try to approach this from a scientific point of view and discuss some of the more commonly used preprocessing techniques in this section.

III Preprocessing

In document clustering, preprocessing include everything from the basic task of converting the indexes into a suitable data representation (e.g. a term-document matrix) to more advanced techniques such as various kinds of parsing the xml document, tokenization, stop word removal, stemming and term weighting.

1) Parsing the XML document

All the markup tags are removed to parse the documents using a parser [19] to take the information inside the body tag into a new file.

2) Tokenization

The text corpus as seen in the screenshot above after parsing is cumbersome and has to be tokenized. Tokenization is the process of breaking parsed document text into chunks, called tokens. This process includes removing the punctuations and the text is lowercased.

3) Stop words Removal

After tokenization, it is needed to remove stop words from the list of words. Stop words like is, are, with, the, from, to etc that occur in almost every document are be removed to proceed further which doesn't provide any use to for weighted index being so common.

4) Stemming

Stemming refers to the process of reducing terms to their stems or root variants. For example agreed-> agree; meetings, meeting -> meet; engineering, engineered, engineer -> engine etc. Stemming reduces the computing time as different form of words is stemmed to form a single word. The most popular stemmer in English is Martin Porter's Stemming Algorithm as shown to be effective in many cases in.

5) Building Inverted Index

Indexing is nothing but refinement i.e. a sufficient general description of a document such that it can be retrieved with a query that contains the same subject as the document and vice versa. Indexing is a mechanism to locate a given query term in a document. Inverted file contains an inverted file entry that stores a list of pointers to all occurrences of that term in the main test for every term in the lexicon, where each pointer is, in effect, the number of a document in which the term appears. There are two types of inverted index. A record level inverted index consists of a list of references to documents for each term. But for further processing we need significant terms that are obtained from dimensionality reduction. This is a major difficulty in text categorization of feature space i.e. total number of terms considered. Even a moderate size collection consists of thousands of unique terms. So we need to reduce the number of terms in the collection which is done by dimensionality reduction. The table 1 shows the inverted file for text.

Number	Term	Documents
1	Cold	1,4
2	Days	3,6
3	Hot	2,3
4	Like	4,5

Once significant terms are obtained, the next step is to find the term frequency and document frequency in order to form vectors for processing clustering algorithm.

6) TF * IDF Calculation

Term Frequency and Inverse Document Frequency is a weight often used in text mining and information retrieval. It is a measure of how important a word is to a document in a collection. Term Frequency is defined as the total count of word that is repeated in a document. Inverse Document Frequency is defined as the total number of times the word occurs in the entire documents i.e. number of documents containing the significant word. Thus the term frequency is given by

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

Here, |D| is the total number of documents in the corpus, $|\{d : t_i \in d\}|$ is the number of documents where the term ti appears (that is ni,j is not equal to 0. If the term is not in the corpus, this will lead to a division by zero. Therefore it is common to use $1 + |\{d : t_i \in d\}|$, Then we define TF-IDF given by

$$(\text{tf-idf})_{i,j} = \text{tf}_{i,j} \times \text{idf}_i \qquad (3)$$

TF * IDF matrix, which representing a vector space model formed by documents that given as input to clustering algorithm where each row represents vector or document and columns show the dimensions of that vector. for example the following table 2 shows the tf*idf matrix,

	Term1	Term2	Term3	Term4	Term5
D1					
D2					
D3					
D4					
D5					

IV Clustering Algorithms

Cluster analysis organises data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups. It has many applications in different areas of computer sciences such as information retrieval, document clustering, computational biology, machine learning, data mining and pattern recognition. Document clustering can be defined as the automatic discovery of document clusters/groups in a document collection, where the formed clusters have a high degree of association (with regard to a given similarity measure) between members, whereas members from different clusters have a low degree of association. The aim of a good document clustering scheme is to minimise intra-cluster distances between documents, while maximising inter-cluster distances. A distance measure thus lies at the heart of document clustering. Several ways for measuring the similarity between two documents exist, some are based on the vector model (e.g. Cosine distance or Euclidean distance) while others are based on the Boolean model (e.g. size of intersection between document term sets). More advanced approaches exist, for instance using Latent Semantic Analysis (LSA) to transform the vector space into a space of reduced dimensionality. There are various clustering algorithms are proposed in the literature like k-means, k-medoids, fuzzy C-Means etc., Here k-means clustering Algorithm is presented for the study.

K-Means Clustering Algorithm

K- Means clustering algorithm was developed by J. MacQueen and then by J. A. Hartigan and M. A.Wong around . K-means is the simplest and most popular classical clustering method that is easy to implement. K-means clustering is an algorithm to classify or to group your objects based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. It is also called centroid method. This method uses the Euclidean distance measure, which appears to work well with compact clusters. The K-means method involves the following steps.

Algorithm:

```
Step 1.  Select the number of clusters. Let this number be k.
Step 2.  Select k seeds as centroids of the k clusters. The seeds may be selected randomly.
Step 3.  Compute the Euclidean distance of each object in the dataset from each of the centroids.
Step 4.  Allocate each object the cluster it is nearest to based
         on the distances computed in the step 3.
Step 5.  Compute the centroids of the clusters by computing
         the means of the attribute values of the objects in each cluster.
Step 6.  Check if the stopping criterion has been met. If yes
         go to step 7, else go to Step 3

Step 7.  One may decide to stop at this stage or to split a
         cluster or combine two clusters heuristically until a
         Stopping criterion is met.
```

Though k-means is simple to implement and provide results, the clusters formed failed for new distance metrics. It fails when the documents size is too large and takes lot of time to run for few metrics.

Dataset

The Text data is available from the publicly available like 20-Newgroups data and reuters document collection. The original data was preprocessed to strip the news messages from the email headers and special tags and eliminate the stop words and stem words to their root forms. Then the words were sorted on the inverse document frequency (IDF), and some words were removed if the idf values were too small or too large. The K-means clustering algorithm is applied to group the documents.

Conclusion and Future Work

In this paper, we concentrate on clustering electronic documents into groups containing similar documents together, based on the clusters formed. In this paper, k-means clustering algorithm is used over the documents after preprocessing. In information reretrival, the process of manually categorising the pages of an electronic / website document is often tedious and expensive. Document clustering has thus often been used to automatically categorise a search result into topic groups (clusters).Other clustering algorithms with different dataset may be applied for performance benchmark as a future work.

References

[1] A. El-Hamdouchi and P. Willet, Comparison of Hierarchic Agglomerative Clustering Methods for
Document Retrieval, The Computer Journal, Vol. 32, No. 3, 1989.

[2] Cai.D,He.X, and Han.J, "Document Clustering Using Locality Preserving Indexing," IEEE Trans. Knowledge and Data Eng.,Vol.17,no.12, Dec.2005.

[3] Gerald Kowalski, Information Retrieval Systems – Theory and Implementation, Kluwer Academic
Publishers, 1997.

[4] Jiawei Han, Micheline Kamber, "Data Mining concepts and Techniques", Morgan Kaufmann Publishers, San Fracisco, CA, USA.

[5] J.Hyma, Y.Jhansi and S.Anuradha, A new hybridized approach of PSO & GA for document clustering, International Journal of Engineering Science and Technology Vol.2(5),2010,1221-1226.

[6] Margaret H. Dunham, Data Mining Introductory and Advanced Topics, Pearson Education in South Asia.

[7] Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, IEEE Press & John Wiley, November 2002.

[8] Michael J.A.Berry Gordon Linoff, Mastering Data Mining" John wiley & sons ptd, Ltd, Singapore 2001.

[9] P.Prabhu and N.Anbazhagan, "Improving the performance of k-means clustering for high dimensional dataset', International Journal of Computer Science and Engineering", Vol 3. No.6. Pg 2317-2322 June 2011.

[10] P.Prabhu,' Discovery of Novel Patterns in Animal Dataset using Hierarchical Techniques', Indian Streams Research Journal, Vol I, Issue V, [June 2011] Information Technology.

Authors Profile

P.Prabhu is working as Assistant Professor in Information Technology, Directorate of Distance Education, Alagappa University, Karaikudi, Tamilnadu, India. He received Master's Degree in Computer Applications from Bharathiar University, Coimbatore in 1993, and M.Phil degree in Computer Science from Bharthidasan University, Trichirappalli in 2005.He has published many articles in National and International Journals. He has presented papers in National and International Conferences. His research area is Data Mining, Information Retrieval and Computer Networks.