DOCUMENT RETRIEVAL USING CLUSTERING

Sivaranjani B1, Yamini C2, Jackulin Durairani A3, Nivi A N4

1PG Scholar, Department of Computer Science and Engineering, Dr. N.G.P Institute of Technology, Anna University, Tamil Nadu, India
2Assistant Professor, Department of Computer Science and Engineering, Dr. N.G.P Institute of Technology, Anna University, Tamil Nadu, India
3PG Scholar, Department of Computer Science and Engineering, Dr. N.G.P Institute of Technology, Anna University, Tamil Nadu, India
4PG Scholar, Department of Computer Science and Engineering, Dr. N.G.P Institute of Technology, Anna University, Tamil Nadu, India

Abstract
The exponential growth of knowledge in the World Wide Web, has understood the need to develop economical and effective ways for organizing relevant contents. In the field of web computing, document clustering plays a vital role and plays an interesting and challenging problem. Document clustering is mainly used for grouping the similar documents in the search engine. The web also has rich and dynamic collection of hyperlink information. The retrieval of relevant document from the internet is the complicated task. Based on the user’s query the document will be retrieved from the various databases to give relevant information and additional information for the given query. The documents are already clustered based on keyword extraction and stored in the database. The probabilistic relational approach for web document clustering is to find the relation between two linked pages and to define a relational clustering algorithm based on probabilistic graph representation. In document clustering, both content information and hyperlink structure of web page are considered and document is viewed as a semantic units. It also provides additional information to the user.

Keywords: Document Clustering, Agglomerative Clustering, Entropy, F-Measure

1. INTRODUCTION
Data mining refers to extracting or mining information from massive databases. Data mining and knowledge discovery in the databases is a new disciplinary field, the statistics, machine learning, databases and parallel computing is used for merging these ideas. Data mining is that the non-trivial method of characteristic valid, novel, probably helpful and ultimately comprehensible patterns in knowledge. The actual data processing task is that the automatic or semi-automatic analysis of huge quantities of knowledge to extract antecedent unknown attention-grabbing patterns like teams of knowledge records, uncommon records and dependencies. This sometimes involves exploitation info techniques such as spatial indexes. With the widespread use of databases and the explosive growth in their sizes, organizations are faced with the problem of information overload. The major problem in all enterprise is effectively utilizing these massive volumes of data.

Clustering is that the method of grouping a group of physical abstract objects into categories of comparable objects. Cluster could be a assortment of knowledge objects that are kind of like alternative inside a same cluster and are dissimilar to the objects in other clusters. In order to enhance the classification task clustering is used as a method to extract information from the unlabelled data. From the unlabelled data cluster is mainly used to create a training set. Technology has been improved a lot in World Wide Web. The increasing size and dynamic content of the World Wide Web has created a need for automated organization of web-pages. Document clusters can provide a structure for organizing large bodies of text for efficient browsing and searching.

Web document clustering has become an important task in analyzing large number of documents distributed among various sites. The main challenge in this clustering method is to organize the documents and produce the better results without introducing much cost and complexity. The retrieval of relevant document from the Internet is the complicated task. Web document clustering discovers useful information from web contents such as text, images, audio, video, metadata and hyperlinks. The web consists not only of pages but also of hyperlinks pointing from one page to another.

The following figure shows how the documents are retrieved from various databases through the search engine.
2. APPROACHES OF THE SYSTEM

2.1. Preprocessing

Raw data is extremely vulnerable to noise, missing values and inconsistency. The standard of data affects the data mining results. To improve the standard of the information and consequently of the mining results is pre-processed thus on improve the potency and simple the mining method. Information preprocessing is one among the foremost critical steps in a data mining process that deals with the preparation and transformation of the initial dataset. The two methods used for preprocessing the given documents are:

- Stop words Removal
- Stemming

The stop words removal approach is used to eliminate the unwanted words such as before, is, a, an, the, become, then, they, there, that, them, etc.

The stemming algorithm is used to eliminate the stemming words and to identify the root words. The stemming words which are ending with ed, ion, ing.

2.2 Retrieving Document from Different Databases

In this module we are going to retrieve all the documents which are relevant to the user given query. The all relevant databases are combined using agglomerative algorithm. When the user submits a query all the relevant links are displayed in the web page. On selecting the particular link the information related to the query are stored in the database.

2.3. Clustering the Relevant Documents

The relevant documents which are stored in the database are clustered based on the given queries. For example, if the user searches the topic “data mining” the documents which are related to the data mining and also additional documents for the given topic are retrieved. All retrieved documents are clustered by considering similarities and combined using agglomerative clustering algorithm.

Manwar et al. developed the vector space or probabilistic model, has term frequency and inverse frequency measures for retrieving documents relevantly. Inter document characterization and document frequency plays a vital role in building ranks of the documents in vector space model.

2.4. Performance Analysis

The clustering performance in terms of effectiveness is measured by two external evaluation metrics.

- F-Measure
- Entropy

3. AGGLOMERATIVE HIERARCHICAL CLUSTERING (AHC)

This bottom up strategy starts by inserting every object in its own cluster and so merges these atomic clusters into larger and bigger clusters, till all of the objects square measure in a very single cluster or till sure termination conditions square measure satisfied. Most gradable bunch strategies belong to the present class. They take issue solely in their definition of entomb clusters similarity. Initially, AGNES places every object into a cluster of its own. The clusters square measure then incorporate in small stages consistent with some criterion. For instance, clusters C1 associated C2 could also be incorporate if associate object in C1 and an object in C2 from the minimum Euclidean distance between any 2 objects from completely different clusters.

This is a single-linkage approach in this every cluster is portrayed by all of the objects within the cluster, and also the similarity between two clusters is measured by the similarity of the closet clusters composition. The cluster merging method repeats till all of the objects square measure eventually incorporates to make one cluster.

Parul Agarwal et al. (2010) established the collective class-conscious technique that works on bottom up approach. the overall approach of class-conscious cluster is in victimization associate applicable metric that measures distance between a pair of tuples and linkage criteria that specifies the unfamiliarity of sets as a operate of the pair wise distances of observations within the sets.

The simple procedure for agglomerative clustering is

- Initially, place every article in its own cluster.
- Among all current clusters, choose the two clusters with the tiniest distance.
- Replace these two clusters with a replacement cluster, shaped by merging the two original ones.
- Repeat the higher than two steps till there’s just one remaining cluster within the pool.
- The result’s a cluster tree. We will cut the tree at any level to provide totally different clump.
4. DATASET AND EVALUATION MEASURE

In order to evaluate the effectiveness of the proposed clustering solution, we compare the performance obtained by our agglomerative clustering. The experimental investigation starts from a dataset construction step, in which about 10,000 web pages from popular sites listed in five categories of Yahoo! Directories (http://dir.yahoo.com/) are downloaded. In order to evaluate and monitor the performance of different clustering methods, we apply a feature selection procedure based on the Term Frequency Variance index.

The clustering performance in terms of effectiveness is measured by three external evaluation metrics, F-Measure, Entropy comparing class labels with cluster assignments. In this experimental investigation we set K = 5, i.e. the number of clusters obtained by all tested algorithms is equal to the number of data set categories.

The F-Measure metric represents a mixture of the exactness and recall live typical of data retrieval.

\[
P(1,j) = \frac{n_{ij}}{n_1} \quad \text{(Eq 1)}
\]

\[
R(1,j) = \frac{n_{ij}}{n_j} \quad \text{(Eq 2)}
\]

Where \(n_{ij}\) is the number of elements and \(n_j\) represents the cardinality of cluster \(j\).

F-Measure

\[
F(1,j) = \frac{2 \times \text{Recall}(1,j) \times \text{Precision}(1,j)}{\text{Precision}(1,j) + \text{Recall}(1,j)} \quad \text{(Eq 3)}
\]

5. EXPERIMENTAL RESULTS

The proposed algorithm has been evaluated comparing its F-Measure, Entropy and Corrected Rand Coefficient, to that one obtained by the k-Means and Expectation Maximization algorithms.

The document collection \(Q\) is submitted to the three algorithms. A pre-processing activity on \(Q\) is performed so as to form page cases insensitive, take away stop words, acronyms, non-alphanumeric characters, hypertext mark-up language tags and apply stemming rules, exploitation Porter’s suffix removal formula. Then, \(Q\) is mapped into a matrix \(M\) where every row of \(M\) represents a document \(d\), following the Vector Space Model.

\[
D_i = [w_{i1}, w_{i2}, ..., w_{i|Z|}] \quad \text{(Eq 4)}
\]

where \(|Z|\) is the number of distinct terms contained in the document set \(Q\) and \(w_{ij}\) is the weight of the \(j\)th term in the \(i\)th document. This weight is computed using the scoring technique TFxIDF, as follows:

\[
W_{ij} = TF(t_j, d_i) \times IDF(t_j) \quad j = 1, ..., |Z|; i = 1, ..., |Q| \quad \text{------- (Eq 5)}
\]

where TF\((t_j, d_i)\) is the Term Frequency, i.e. the number of occurrences of term \(t_j\) in \(d_i\), and IDF\((t_j)\) is the Inverse Document Frequency.

IDF\((t_j)\) is a factor which enhances the terms which appear in fewer documents, while downgrading the terms occurring in many documents and is defined as

\[
IDF(t_j) = \log(|Q|/DF(t_j)) \quad j = 1, ..., |Z| \quad \text{------- (Eq 6)}
\]

where \(DF(t_j)\) is that the variety of documents containing the \(j\)th term.

Moreover, since all the compared approaches depend on the initial choice of the representative element of each cluster (centroids) during the initialization phase, we report the obtained performance over 500 runs. In particular we show, both for Entropy, F-Measure and Corrected Rand constant, their individual minimum, maximum, average value and confidence interval (confidence level at 95%).

This means that the performance have been evaluated on a vocabulary dimensioned as \(T = 20; 50; 100\)

| Table 1: Performance comparison with 20, 50 and 100 Terms |
|---------------|---------------|---------------|------------------|
| F-Measure | Min | Max | Average | Confidence Interval |
| 20 EM | 0.443 | 0.738 | 0.629 | 0.602-0.630 |
| KM | 0.38 | 0.708 | 0.551 | 0.548-0.557 |
| AGG | 0.445 | 0.799 | 0.659 | 0.665-0.663 |
| 50 EM | 0.445 | 0.855 | 0.669 | 0.662-0.672 |
| KM | 0.455 | 0.788 | 0.629 | 0.625-0.638 |
| AGG | 0.573 | 0.875 | 0.748 | 0.724-0.751 |
| 100 EM | 0.432 | 0.931 | 0.936 | 0.679-0.691 |
| KM | 0.511 | 0.819 | 0.819 | 0.682-0.694 |
| AGG | 0.585 | 0.899 | 0.899 | 0.752-0.761 |
Table 2: Performance comparison with 20, 50 and 100 Terms

<table>
<thead>
<tr>
<th>Entropy</th>
<th>Min</th>
<th>Max</th>
<th>Average</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>EM</td>
<td>0.639</td>
<td>1.281</td>
<td>0.862,0.851-0.869</td>
</tr>
<tr>
<td></td>
<td>KM</td>
<td>0.759</td>
<td>1.429</td>
<td>1.042,1.031-1.050</td>
</tr>
<tr>
<td></td>
<td>AGG</td>
<td>0.462</td>
<td>1.069</td>
<td>0.659,0.650-0.669</td>
</tr>
<tr>
<td>50</td>
<td>EM</td>
<td>0.399</td>
<td>1.338</td>
<td>0.762,0.749-0.773</td>
</tr>
<tr>
<td></td>
<td>KM</td>
<td>0.583</td>
<td>1.229</td>
<td>0.872,0.863-0.885</td>
</tr>
<tr>
<td></td>
<td>AGG</td>
<td>0.342</td>
<td>0.862</td>
<td>0.503,0.499-0.513</td>
</tr>
<tr>
<td>100</td>
<td>EM</td>
<td>0.209</td>
<td>1.179</td>
<td>0.672,0.658-0.685</td>
</tr>
<tr>
<td></td>
<td>KM</td>
<td>0.452</td>
<td>1.075</td>
<td>0.734,0.729-0.745</td>
</tr>
<tr>
<td></td>
<td>AGG</td>
<td>0.392</td>
<td>1.027</td>
<td>0.752,0.863-0.885</td>
</tr>
</tbody>
</table>

6. CONCLUSION

The Conclusion of the web document clustering is that clustering is a very useful technique to deal with a large, heterogeneous and dynamic web page collections efficiently. The indexing and retrieval will be optimized once the documents are clustered together in a sensible order. It improves the quality of the data and also improves the efficiency of the mining process. The documents are clustered based on keyword extraction. By using Expectation Maximization and the agglomerative algorithm we can effectively extract topics contained in the different documents.
REFERENCES


