# Enhancing Document Clustering through Heuristics and Summary-Based Pre-processing

Sri Harsha Allamraju and Robert Chun

San Jose State University, Department of Computer Science, San Jose, CA 95192
sriharsha451@gmail.com, Robert.Chun@sjsu.edu

**Abstract.** Knowledge workers are burdened with information overload. The information they need might be scattered in many places, buried in a file system, in their email, or on the web. Traditional Clustering algorithms help in assimilating these wide sources of information and generating meaningful relationships amongst them. A typical clustering preprocessing involves tokenization, removal of stop words, stemming, pruning etc. In this paper, we propose the use of summary and heuristics of a document as a pre-processing technique. This technique preserves the formatting of a document and uses this information for producing better clusters. In addition, only a summary of a document is used as the basis for clustering instead of the whole document. Clustering algorithms using the proposed pre-processing technique on formatted documents resulted in improved and more meaningful clusters.

**Keywords:** document clustering, clustering, summarization, heuristics.

## 1 Introduction

In today's information age, a typical computer user's information is stored in many places. This information is stored in many forms but can be broadly classified into two forms, namely formatted and un-formatted documents. Unformatted documents typically include plain-text files, whereas formatted documents include word documents, presentations, and web pages. File managers allow users to store information in tree-structured hierarchies, also known as folders. Thus the user faces the ontological burden of classifying one's documents and storing them in relevant folders. With hard disk storage space in the orders of gigabytes, the user is burdened with information overload.

Clustering algorithms help in grouping objects into one cluster based on a similarity measure. When applied on documents, this can help in grouping similar documents based on content. Many clustering algorithms have been proposed for clustering documents. A typical clustering process involves tokenization, removal of stop words, stemming and pruning.

A formatted document such as a word document typically consists of headings, emphasized words, de-emphasized words, italicized words etc. Also the font, size and color of text in these documents vary. This implies that some words in the document are more important than others. This importance is evident from the increased human readability of a formatted document over that of an unformatted one.

M.J. Smith and G. Salvendy (Eds.): Human Interface, Part II, HCII 2009, LNCS 5618, pp. 105–113, 2009.

In the traditional clustering pre-processing step, the document is first tokenized. However, once tokenized, the words lose their formatting. This implies that all words contribute equally to clustering irrespective of its formatting in the original document. In this paper, we propose an additional pre-processing step. For each document, a representative document is generated. This is obtained by combining the summary of a document with its heuristics. The heuristics of a document is the set of words which are emphasized in the document through the author's use of various formatting techniques. For example, words that are bolded, underlined, italicized, or that appear in headings contribute valuable heuristics concerning the document's content. The summary of a document is obtained by using a document summarization algorithm. The document's summary, together with its heuristics, is clustered instead of the whole document.

The proposed approach has many advantages. Firstly, it takes into account the formatting of the document. This helps in identifying words in the document which are more important and representative of the document's content. Secondly, only the summary of the document is utilized for clustering rather than the entire document. This helps in reducing the "noise" in a document and gathers sentences that are of utmost importance. Therefore, the summary of the document presents a realistic view of the document's content. Thirdly, the proposed pre-processing helps in producing more accurate and realistic clusters.

The rest of the paper is organized as follows. Section 2 describes existing work done with relevance to the topic of this paper. Section 3 consists of a detailed explanation of the proposed pre-processing technique. Section 4 provides experimental results. Section 5 outlines some concluding remarks, ongoing research and future direction of work.

## 2 Related Work

Clustering algorithms have a variety of applications and are used in various fields such as image segmentation, object recognition and information retrieval. Different Clustering algorithms work best for different types of data. Jain et al. [1] provides an overview of Data Clustering techniques.

Budzik et al. [2] proposed a system which extracts keywords from a document that are representative of the document's content. These keywords are later fed to a web search engine and web pages related to the context in which the user is working is shown. In order to extract search terms from a document, the authors proposed a set of heuristics. A subset of these heuristics forms the basis of the proposed pre-processing technique discussed here. CACTUS [9] attempts to cluster categorical data using summaries.

Visser et al. [3] built an automatic summarizer system based on word frequency count, cue phrase, location, title and query method. In the word frequency method, each sentence is assigned a score based on the relevant words in that sentence. In a cue phrase method, each sentence was assigned a cue score based on the presence of relevant and important phrases. In the location method, a score is assigned to the sentence based on its location in a paragraph or proximity to headings. In the title method, sentences containing words present in the document's title are given a higher score. In the query method, sentences matching the query words are given more importance. The final score

of each sentence is obtained by weighted sum of above-mentioned features. Thus, the summary obtained gives the list of sentences which are of utmost importance and most representative of the document's content. This summarization technique is also used as a basis of the proposed pre-processing technique described next.

## 3 Proposed Pre-processing Technique

Traditional Clustering algorithms pre-process the input data through tokenization, stemming, pruning and removal of stop words. This works well for unformatted documents such as plain-text files. However by using the same pre-processing technique for formatted documents such as word documents, presentations, web pages etc., certain important information is lost. This is reflected in the quality of clusters thus obtained. This paper stands by the premise that formatted documents contain some more information than plain-text files and so should be treated differently. The proposed pre-processing step consists of two components, namely document heuristics and summarization. The following section describes each of them in detail.

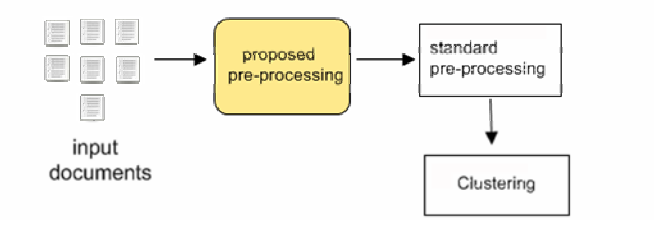


**Fig. 1.** Positioning of proposed pre-processing technique with respect to existing practice

### 3.1 Document Heuristics

Budzik et al. [2] proposed a set of heuristics for extracting important keywords from a document. They are as follows: 1) remove stop words, 2) value frequently used words, 3) value emphasized words, 4) value words that appear in the beginning of the document rather than at the end, 5) punish words appearing to be intentionally de-emphasized, 6) ignore the ordering of words in a list and 7) ignore words that occur in sections of the document that are not indicative of the document's content.

Of the heuristics mentioned above, some of the heuristics are not applied here since they are taken into consideration by the document summarizer. One of the main heuristics used by the proposed pre-processing technique is to value emphasized words and to punish words appearing to be intentionally de-emphasized. Emphasized words refer to the set of words that are formatted as bold characters, italicized, underlined, appear in capital letters and in headings. In addition, words that are colored are also considered emphasized. De-emphasized words refer to words that have a font size smaller than that of the majority of words in the document. Thus, by subjecting a document to heuristics, a set of emphasized words is obtained. However, in the case of plain-text files, there is no formatting present. Therefore, these particular heuristics will not produce any results on plain text documents.

### 3.2 Document Summary

Automatic Text Summarization is one of the important aspects of the proposed pre-processing technique. The main idea of using automatic summarization is to use the portion of the document that is most important and that can represent the whole document in terms of its content and context. By using automatic summarization, only those sentences of the document are obtained which are most relevant. Thus, this reduces the noise in the clustering data that might be obtained due to the presence of unwanted sentences and words. Therefore, instead of performing clustering on the whole document, only the summary and heuristics of the document are used.

The automatic summarizer built by Visser et al. [3] was created for generating summaries of scientific documents. Since most of the scientific documents such as research papers are well formatted, the following summarizer was chosen for improved accuracy. The summary is generated by the system based on weighted scores obtained by word frequency count, cue phrase, location, title and query. However, since there is no query involved in clustering, a weight of zero is assigned for the query method. In addition, the original summarizer was designed to work more effectively for scientific documents. The cue phrase method looks for certain phrases most frequently found in research papers and other scientific documents. To keep the summarizer more generic in nature, the cue phrase method is also assigned a weight of zero.

### 3.3 Process

The whole process of the proposed pre-processing technique is as shown in Fig 2. The document is first sent to a heuristic analyzer. It returns a set of emphasized words in the document. Then, the document is fed to an automatic summarizer. This returns the summary of the document. The summary along with the words obtained from heuristics is stored in a file. This new file is a representative document for the original formatted document. Thus for each document in the corpus which is to be clustered, a representative document is generated which consists of its summary and heuristics. These representative documents are used for clustering instead of the original documents. The standard pre-processing techniques are still applied to the representative documents and then sent as input to the clustering engine. Once the representative documents are clustered, they are mapped back to their original document.

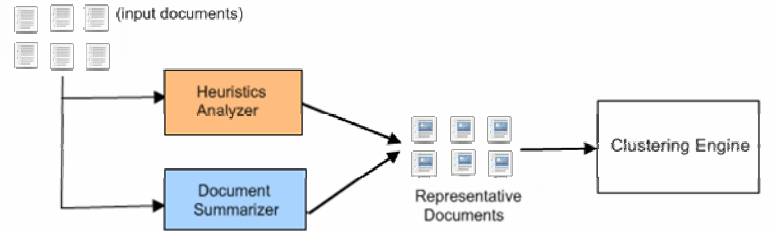


**Fig. 2.** Proposed Pre-processing technique in detail

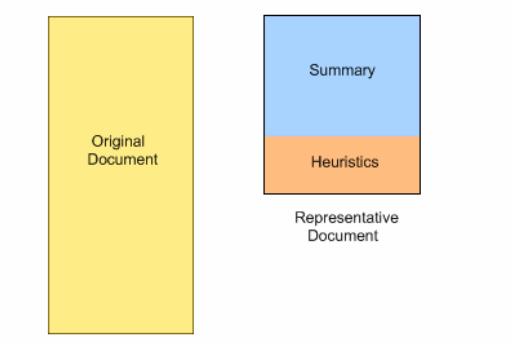


**Fig. 3.** Comparison between original document and its representative document

# 4 Experimental Setup

## 4.1 Clustering Engine

The CLUTO [4] clustering toolkit was used for clustering documents. It is a highly scalable toolkit and can be used on high dimensional dataset. A wide variety of clustering algorithms such as partitional, agglomerative and graph-partitioning based have been implemented in this toolkit. An extensive selection of similarity measure functions is available such as Euclidean, Cosine, correlation co-efficient and Jaccard. In addition, a user defined similarity measure can be used.

The CLUTO clustering toolkit provides detailed reports for each clustering activity. Different external quality measures such as entropy and purity are computed for each cluster. In addition, CLUTO implements five new clustering criterion functions proposed by Karypis et al. [5].

## 4.2 Visualization Engine

In order to understand the topology of the clusters that are obtained, gCLUTO [6] was used for visualization. It produces two types of visualizations on cluster data, namely Matrix Visualization and Mountain Visualization. The Mountain Visualization was used here for analyzing the results.

The Mountain Visualization technique uses peaks to represent clusters. The degree of separation of one cluster from the other denotes the relative similarity of clusters. Clusters that are very similar to each other are close to each other and in some cases overlap. This visualization is effective in understanding the relationships between clusters.

Each peak corresponds to a single cluster. These peaks are of varying height, size and color. The height of the peak represents the internal cluster similarity. The higher the peak the greater is the internal similarity and vice versa. The volume of the peak represents the number of documents in a cluster. The color of the peaks represents the internal standard deviation. Different colors mean different levels of deviation. Red represents a low standard deviation, whereas blue corresponds to high

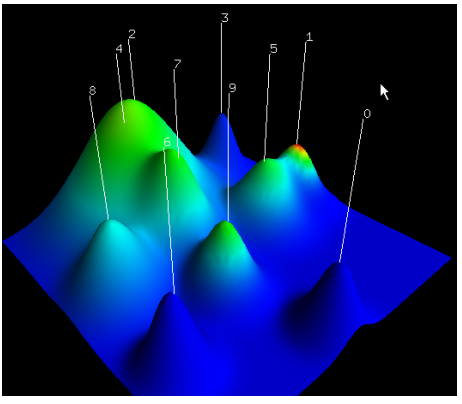


**Fig. 4.** Mountain Visualization of sample data showing peaks of different heights and colors

standard deviation. Clusters with high standard deviations are noisy and are definitely unwanted. Colors such as red, orange, yellow and green represent standard deviations from low to medium range, with red representing least standard deviation.

Thus an ideal clustering would be expected to have distinct clusters with low standard deviation and high internal cluster similarity. This corresponds to a mountain visualization consisting of high non-overlapping peaks colored in red or orange.

### 4.3 Clustering Algorithms

Different clustering algorithms work best for different datasets. Primarily, the agglomerative clustering algorithm was applied on documents for clustering. Given the required number of clusters, the agglomerative algorithm first assigns each document to its own cluster and then merges other documents repeatedly until the required number of clusters is obtained. The criterion used for merging one document into a cluster depends on the merging schemes. The CLUTO clustering toolkit supports a wide variety of merging schemes, namely single-link, complete-link and group average approaches in addition to seven new merging schemes.

### 4.4 Dataset

The Reuters Transcribed Subset dataset[1] was used to evaluate the effectiveness of the proposed pre-processing technique. This dataset is a subset of Reuters-21578 collection[2]. The Reuters Transcribed dataset consists of 20 files picked from each of the 10 largest classes in the Reuters-21578 collection. These files were generated by an automatic speech recognition (ASR) system. This possibly introduces a certain degree of noise in the data. However, the proposed pre-processing technique is not highly effective when applied to plain-text documents without any formatting. Thus, all the 200 files were manually formatted using headings, bolds, italics and other forms of emphasis.

[1] Available at http://kdd.ics.uci.edu/databases/reuters_transcribed/reuters_transcribed.html
[2] Available at http://www.daviddlewis.com/resources/testcollections/reuters21578/

## 5 Results

The Reuters Transcribed dataset is formatted and then subjected to the proposed pre-processing technique. The document is sent to a heuristics analyzer and a document summarizer. The words returned by the heuristic analyzer are merged with the summary obtained from the document summarizer. This merged document becomes the representative document for the original document and is clustered on behalf of the original document.

An agglomerative k-means clustering algorithm is used. Since the original dataset is manually classified into 10 categories, the value of k for the k-means clustering algorithm is given as 10. The CLUTO clustering engine applies the agglomerative k-means clustering algorithm and tries to divide the given input dataset into 10 clusters.

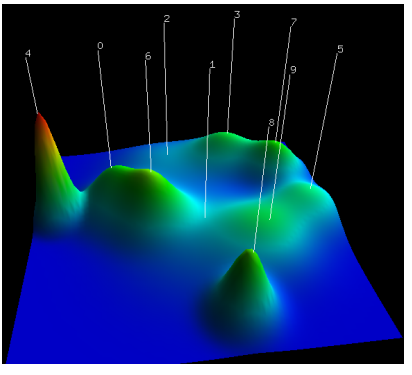


**Fig. 5.** Mountain Visualization of Reuters Transcribed Subset Dataset using standard clustering techniques

Once the clusters are obtained, the gCLUTO software is used to visualize the resulting cluster formation. The gCLUTO software takes a matrix file as input. A matrix file is an intermediate file generated by the CLUTO clustering engine. The columns in the matrix file correspond to unique words in the given document corpus. The rows represent the document number. Once the required matrix file is given as input to gCLUTO, Mountain Visualization of resulting clusters is generated and displayed.

In the first experiment, the test dataset is pre-processed using standard clustering techniques and then sent to the clustering engine for k-means agglomerative clustering. The resultant clusters obtained are as shown in Fig. 5.

By observing the Mountain Visualization (shown in Fig. 4.) it is seen that some of the clusters are overlapped and are not separated such as clusters (0,6) and (3,7). Most of the peaks (clusters) are green in color indicating a comparatively high standard deviation. In addition, cluster 2 is pale blue in color indicating very high standard deviation. Cluster 4 is the highest of all indicating greater internal similarity.

In the second experiment, the test dataset is pre-processed using the proposed pre-processing techniques and then sent to the clustering engine for k-means agglomerative clustering. The resultant clusters obtained are as shown in Fig. 6.

The Mountain Visualization of cluster distribution obtained by applying the proposed summary and heuristics-based pre-processing produced better and elegant results. This is evident from the visualizations obtained as shown in Fig. 6. Most of the clusters are in red, orange or yellow colors indicating a low standard deviation within clusters. Most of the clusters are of similar height indicating a uniform distribution of internal similarity across clusters. Cluster 4 is the highest peak, indicating high internal similarity. The clusters are evenly distributed and well separated from one another.

In comparison with the standard clustering technique, it is observed that the proposed pre-processing technique helped in creating distinctly separated clusters. Also, the overall internal similarity of elements within a cluster is increased. Additionally, by using the proposed heuristics and summary-based pre-processing, there has been an improvement in standard deviation within clusters. The proposed pre-processing produced clusters with low internal standard deviation. It can supplement, and can be used in conjunction with, existing approaches to clustering. The augmentation of traditional clustering techniques with our proposed heuristics and summary based technique does not add significant processing times – the analysis of the 200 News articles took just 1.2 seconds more than otherwise.

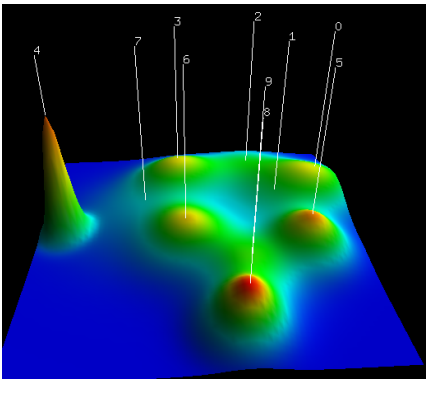


**Fig. 6.** Mountain Visualization of Reuters Transcribed Subset Dataset using proposed pre-processing techniques

## 6 Conclusions and Future Work

In this paper, we proposed summary and heuristics-based pre-processing for document clustering. We have shown that this pre-processing technique results in clusters with improved internal and external cluster quality measures than compared to existing clustering techniques. We have used a visualization technique as the basis for such a conclusion. This states that when clustering formatted documents, using just the summary and heuristics themselves are enough and that the whole text of the document may not be necessary. Heuristics and Summary-based pre-processing opens a new dimension in document clustering.

The proposed pre-processing technique has been applied only on the Reuters Transcribed Subset dataset. The plain-text dataset was converted into individual word documents and the news article title was manually set to bold for each file. However, in case of research papers, which follow certain naming conventions for headings such as “Abstract”, “Related Work”, “Conclusions” etc. it would be interesting to observe if the proposed pre-processing would actually be able to classify them into different classes. In short, the effectiveness of the proposed pre-processing steps should be examined using a wider variety of input files as test data, and that it is early to conclude the generality of the proposed pre-processing technique.

In this paper, the given dataset is pre-processed using the proposed technique and then subjected to agglomerative k-means clustering. However, it would be interesting to see the effect of using other clustering algorithms on the proposed pre-processing technique. In addition, it would be interesting to note the effect of different summary ratios on the quality of clusters. This would lead to finding an ideal summary ratio value where the quality of clusters produced is optimal for this kind of pre-processing. Finally the effect of using a weighted approach to summary and heuristic pre-processing should be investigated.

## References

1. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys 31(3), 264–323 (1999)
2. Budzik, J., Hammond, K.J., Birnbaum, L.: Information access in context. Knowledge-Based Systems 14, 37–53 (2001)
3. Visser, W.T., Wieling, M.B.: Sentence-based Summarization of Scientific Documents. The design and implementation of an online available automatic summarizer. Report (2009), http://home.hccnet.nl/m.b.wieling/files/wielingvisser05automaticsummarization.pdf (last retrieved Febuary 12, 2009)
4. Zhao, Y., Karypis, G.: Evaluation of hierarchical clustering algorithms for documents datasets. In: International Conference on Information and Knowledge Management, McLean, Virginia, United States, pp. 515–524 (2002)
5. Zhao, Y., & Karypis, G.: Criterion functions for document clustering: Experiments and analysis. Technical report, Department of Computer Science, University of Minnesota
6. Rasmussen, M., Karypis, G.: gCLUTO: An interactive clustering, visualization and analysis system. Technical Report 04-021, University of Minnesota, s (2004)
7. Reuters-21578 Dataset, http://kdd.ics.uci.edu/databases/reuters_transcribed/reuters_transcribed.html
8. Reuters Transcribed Subset Dataset, http://www.daviddlewis.com/resources/testcollections/reuters21578/
9. Ganti, V., Gehrke, J., Ramakrishnan, R.: CACTUS-Clustering Categorical Data Using Summaries. In: Proceedings of the ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, United States (1999)