Product Development Analyzer using Document Clustering with Chatbot

Saurabh Khedkar¹, Ayesha Ansari², Saloni Waikar³, D. S. Gogawale⁴
Professor⁴
Department of Computer Engineering
Navsahyadri Education Society’s Group of Institutions, Pune, Maharashtra, India

Abstract:
Descriptive document clustering consists of automatically organizing data instances into clusters and generating a descriptive summary for each cluster. The description should inform a user about the contents of each cluster without further examination of the specific instances, enabling a user to rapidly scan for relevant clusters. Selection of descriptions often relies on heuristic criteria. We have developed model descriptive clustering as an auto-encoder network that predicts features from cluster assignments and predicts cluster assignments from a subset of features. The subset of features used for predicting a cluster serves as its description. For text documents, the occurrence or count of words, phrases, or other attributes provides a sparse feature representation with interpretable feature labels. In the proposed network, cluster predictions are made using logistic regression models, and feature predictions rely on logistic or multinomial regression models. Optimizing these models leads to a completely self-tuned descriptive clustering approach that automatically selects the number of clusters and the number of features for each cluster. We applied the methodology to a variety of short text documents and showed that the selected clustering as evidenced by the selected feature subsets are associated with a meaningful topical organization.

Keywords: Document Clustering, Descriptive clustering, feature selection, Intra Document Clustering, Inter Document Clustering, Sentiment analysis, SentiWordNet, Stop Word Removal, String Matching, K-means.

1. INTRODUCTION

Every day the mass of information available, merely finding the relevant information is not the only task of automatic text classification systems. Instead the automatic text classification systems are assumed to retrieve the relevant information and are organized according to their degree of relevance to the query. The main problem in organizing is to classify which documents are relevant and which are irrelevant. The Automated text classification consists of automatically organizing clustered data. Due to hard language of the paper as well as difficult technical terms so it is tedious job for student to understand and implement the paper. To help student for understanding IEEE papers, product development analyser consist all the guidance regarding IEEE paper. Data analysis techniques such as clustering can be used to identify subsets of data instances with common characteristics. Users can then explore the data by examining some instances from the full dataset. This enables users to efficiently focus on relevant subsets of large datasets, especially for collections of documents. Descriptive clustering consists of automatically grouping sets of similar instances into clusters and automatically generating a human-interpretable description or summary for each cluster. The quality of the clustering is important to provide a user with an informative and concise summary that accurately reflects the contents of the cluster. We consider a direct correspondence between description and prediction. We assume each instance is represented with sparse features (such as a bag of words), and each cluster will be described by a subset of features.

2. SYSTEM ARCHITECTURE

![System Architecture Diagram](image-url)

Figure 1. System Architecture
Above figure shows the proposed framework design. The system includes various modules such as document clustering, chatbot, development lifecycle. In document clustering module, it consist all the guidance regarding IEEE paper that helps students to understand the paper. In development lifecycle module it consist all the guidelines and related links of product development process (Requirement gathering, Requirement Analysis, Designing, Testing, Deployment), that helps student to build the project. And there is also ChatBot module for any queries regarding all the paper or phases.

2.1 DOCUMENT CLUSTERING
Document clustering [1] is a very useful tool in today's world where a lot of documents are stored and retrieved electronically. It enables one to discover hidden similarity and key concepts. Moreover, it enables one to summarize a large amount of documents using key or common attributes of the clusters. Thus clustering can be used to categorize document databases and digital libraries, as well as providing useful summary information of the categories for browsing purposes. Automatic clustering of the documents enables the user to have a clear and easy grasp of what kind of documents are retrieved, providing tremendous help for him/her to locate the right information. Text clustering is known as an unsupervised and automatic grouping of text documents into clusters, so that documents within a cluster have a high similarity between them, but they are dissimilar to documents in other clusters. Any clustering techniques have been developed and they can be applied to clustering documents. [2], [3], [4], [5] contain examples of using such techniques. Most of these traditional approaches use documents as the basis of clustering. In [6] there is an alternative approach, wordset based document clustering (WDC). Where instead of comparing documents and clustering them directly or based on some common words, we cluster the document based on closed wordsets found in such documents. First we search global frequent wordsets and frequent closed wordsets in the documents. For each global frequent wordset, we form an initial cluster containing documents that have that wordset. After that, we make disjoint the clusters that contain similar sets of documents, by a score function approach [7]. The resulting clusters will contain documents that share a similar set of words. Document clustering method consists of the following phases: finding frequent closed wordsets, creating initial clusters for each closed wordset, making clusters disjoint using score function, building cluster tree, and tree pruning. Makin clusters disjoint, assign a document to the best initial cluster. If there are several best clusters, the document is kept only in the cluster identified by longest label (in terms of the number of items). A cluster Ci is good for a document docj if there are many global frequent words in docj that appear in many documents in CL. Assign each docj to the initial cluster Ci that has the highest score as shown in following equation:

\[
\text{Score}(C_i \leftarrow \text{doc}_j) = \left( \sum_x n(x) \cdot \text{cluster\_support}(x) \right) - \left( \sum_{x'} n(x') \cdot \text{cluster\_support}(x') \right)
\]

Where,
- \(x\) represents a global frequent word in \(\text{doc}_j\) and the word is also cluster frequent in \(C_i\).
- \(x'\) represents a global frequent word in \(\text{doc}_j\) but the word is not cluster frequent in \(C_i\).
- \(n(x)\) is the frequency of word \(x\) in the feature vector of \(\text{doc}_j\).

2.2 DESCRIPTIVE CLUSTERING
The motivation for applying descriptive clustering to text datasets is that it can be used as an information retrieval mechanism. A user can efficiently scan the descriptions for relevancy versus having to determine which clusters are relevant by manually checking the document instances. Scatter-gather [10], [11], [12], [13] is an iterative procedure that uses multiple stages of descriptive clustering to help a user find relevant documents. An initial clustering is given along with some description or preview of each cluster to the users, who are then asked to select clusters of interest. Instances within the selected clusters are combined and clustered again. This continues until a user hones in on a relevant set of documents. The quality of the automatic description is crucial to enable a user to recognize which clusters are relevant. Descriptive clustering can be performed by firstly clustering and then finding the set of features associated with each cluster. This enables any applicable clustering algorithms to be used. Selecting features that best inform a user on the contents of a cluster is the subsequent challenge. The most basic approach is to describe each cluster by the most likely words in the cluster [10], titles (if available) near the center of each cluster [10], or phrases with similar context as the most likely words [14]. However, these features may not be optimal for discriminating between different clusters. Other scoring criteria such as mutual information [15] (i.e., information gain [16] rather than point wise mutual information [17]) may be used to select more discriminating features (e.g., keywords or phrases) for the clusters. With the aforementioned approaches it is not clear how to objectively measure the selected feature lists or labels that serve as descriptors. Our hypothesis is that any description is only useful if it would enable a user to accurately predict the contents of the cluster. In this case, finding sets of features to describe each cluster can be seen as a feature selection problem. For instance, one can train a decision tree for each cluster to classify instances directly by the presence or absence of certain features. The boolean expression corresponding to the decision tree serves as a description of the set.

2.3 FEATURE SELECTION
Given a particular cluster, predicting whether instances belong to the cluster is a standard classification problem, and selecting the best subset of features for this task is a feature selection problem [16]. Choosing a small subset of maximally predictive features is a difficult task. Feature selection is a basic step in the construction of a vector space or bag-of-words model. In
particular, when the processing task is to partition a given document collection into clusters of similar documents a choice of good features along with good clustering algorithms is of paramount importance. It is aiming at extract a small subset of feature from the problem domain while retaining a suitably high accuracy in representing the original features. In recent years, a lot of feature selection methods have been proposed. There are some traditional feature selection methods based on Document Frequency (DF) and Term Strength (TS). In [19] a novel feature selection method based on part-of-speech and word co-occurrence. In the traditional process of feature selection in document clustering, researchers segment each document into a list of terms and use some criterions to score and sort these candidate features. In this way, we treat each segmented term equally. However, this approach ignores the information of each term’s part of speech. A document or a sentence is made of a list of full words (noun, verb, adjective…) and functional words (preposition, conjunction, auxiliary…). These functional words do not contain the semantic meanings and only are used to be some syntactic elements, which are meaningless to the whole document in the semantic aspect. Meanwhile, to certain separated term in the whole document set, those words, which occur front or behind this term, hold valuable information to explain its meaning. We call these words Context Words. In other words, one term co-occurs with its context words in a document. Considering the analysis above, we use terms’ context words to select the features and to measure the contribution of certain term to document clustering.

2.4 SENTIMENT ANALYSIS
The World Wide Web contains a huge amount of opinion expressing documents containing comments, feedback, critiques, reviews and blogs, etc. They contain very valuable information for helping people in their decision making. For example, product reviews can help enterprises to promote their products; comments on a policy can help politicians to clarify their political strategy, etc. However, the number of these types of documents is huge and they are usually expressed in natural language, so it is impossible for humans to read and analyze all of them. Thus, work which helps to automatically determine the sentiment direction (positive or negative) of online texts is significant. The task of developing such a technique is often called sentiment analysis or opinion mining. It refers to a broad area of natural language processing, computational linguistics and text mining.

In paper[20], they introduces clustering-based sentiment analysis approach which is a new approach to sentiment analysis. By applying a TF-IDF weighting method, voting mechanism and importing term scores, an acceptable and stable clustering result can be obtained. For representing text documents, we use a bag-of-words representation. Each feature dimension is weighted by the logarithm of the inverse occurrence rate, the standard term-frequency inverse-document frequency (TF-IDF), and instances are normalized to have unit-norm. For clustering text documents, we use spectral clustering applied to the similarity matrix implicitly formed from the cosine similarity between the TF-IDF vectors [21], and the spectral clustering algorithm of Ng et al. [22] is applied to the similarity matrix. All instances are used by the clustering algorithm, but a subset of instances are used as training instances for selecting feature subsets and evaluating the predictive features.

2.5 STOP WORD REMOVAL
The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

What are Stop words?
A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query

2.6 STRING SIMILARITY

<table>
<thead>
<tr>
<th>SPELLING ERRORS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. It’s “calendar”, not “calender”.</td>
</tr>
<tr>
<td>2. It’s “definitely”, not “definitely”.</td>
</tr>
<tr>
<td>3. It’s “tomorrow”, not “tommorrow”.</td>
</tr>
<tr>
<td>4. It’s “noticeable”, not “noticeable”.</td>
</tr>
<tr>
<td>5. It’s “convenient” not “conivient”</td>
</tr>
</tbody>
</table>

Figure 2. Example of String Similarity

Word similarity matching is an essential part for text cleaning or text analysis. Let’s say in your text there are lots of spelling mistakes for any proper nouns like name, place etc. and you need to convert all similar names or places in a standard form.

2.7 K-MEANS CLUSTERING
Clustering is an unsupervised learning mode, to draw inferences from datasets consisting of input data without class label or target values, or for exploratory data analysis to find hidden patterns. It groups data instances that are similar to each other in a cluster and data instances that are very dissimilar from each other into different clusters. Cluster analysis is about forming the clusters or organizing the data such that the intracluster distance is less and intercluster distance is more. There are popular algorithms to prepare clusters. In this paper we use k-means clustering algorithm. K-means is the most important flat clustering algorithm. The objective function of K-means is to minimize the average squared distance of objects from their cluster center, where a cluster center is defined as the mean or centroid μ of the objects in a cluster C:

$$\mu(C) = \frac{1}{|C|} \sum_{x \in C} x$$

The ideal cluster in K-means is a sphere with the centroid as its center of gravity. Ideally, the clusters should not overlap. A measure of how well the centroids represent the members of their clusters is the Residual Sum of Squares (RSS), the squared distance of each vector from its centroid summed over all vectors

$$RSS_i = \sum_{x \in C_i} ||x - \mu(C_i)||^2$$

$$RSS = \sum_{i=1}^{K} RSS_i$$
K-means can start with selecting as initial clusters centers K randomly chosen objects, namely the seeds. It then moves the cluster centers around in space in order to minimize RSS. This is done iteratively by repeating two steps until a stopping criterion is met

1. Re-assigning objects to the cluster with closest centroid.
2. Re-computing each centroid based on the current members of its cluster.

### Demonstration of the standard algorithm

1. \( k \) initial "means" (in this case \( k = 3 \)) are randomly generated within the data domain (shown in color).
2. \( k \) clusters are created by associating every observation with the nearest mean.
3. The centroid of each of the \( k \) clusters becomes the new mean.
4. Steps 2 and 3 are repeated until convergence has been reached.

### 3. CONCLUSION

We develop product in physical form by implementing our theoretical knowledge and no doubt it’s a great experience for us. If we want to decrease the uncertainty of new product introductions into the marketplace and to maximize the return on investment of successful launches, it is wise to use market research methods developed to help marketing and product managers to predict the likelihood of a new product's acceptance in the marketplace. Obviously, we should go through field test to ensure its authenticity to meet challenges in real situation. In which it will be ultimately subjected and a positive corporate mind-set is also imperative to be successful in new product development. Encouragement and rewards for innovative concepts and ideas should be used as an incentive for all staff involved in the process, not just the marketing and product managers. The overall corporate philosophy on new product development should view the process as part of the investment strategy and the main focus will be on fitting and feeling of the product as we believe the whole success or failure depend on it.

### 4. REFERENCES


[8]. Sun Park, Dong Un An, Choi Im Cheon “Document Clustering Method using Weighted Semantic Features and Cluster Similarity”.


[20]. Gang Li, Fei Liu “A Clustering-based Approach on Sentiment Analysis” Department of Computer Science and Computer.
