Application of Clustering Algorithms to Group Medical Documents

Ravi Seeta Sireesha
Research Scholar
Department of Computer Science and Systems Engineering
Andhra University College of Engineering
Andhra University
Visakhapatnam

P. S. Avadhani
Professor
Department of Computer Science and Systems Engineering
Andhra University College of Engineering
Andhra University
Visakhapatnam

ABSTRACT
Medical documents contain valuable information about medication and symptoms, which help in improving health care. Recently, large volumes of medical documents are generated by electronic health record systems. These medical documents are unstructured or semi-structured from which extraction of useful information is a difficult task. Application of document clustering techniques is an efficient way for navigation and summarization of documents and very important for many natural language technologies [1]. Various partitional and agglomerative clustering techniques are applied in order to cluster the medical documents for grouping them into meaningful clusters.

General Terms
Medical documents, unstructured documents.

Keywords
Partitional and Agglomerative Clustering techniques, summarization of documents.

1. INTRODUCTION
Medical documents contain lot of valuable information about patients, such as medication conditions (diseases, injuries, medical symptoms, etc.) and responses (diagnoses, procedures, and drugs) [2]. These resources have a huge potential to improve health care which are underutilized This valuable information extracted from clinical notes can be used to build profiles for individual diseases, syndromes, discover disease correlations and enhance patient health care. Medication information in such documents contains information about the medication names, dosage, frequency and duration of drugs [1]. Medical documents are unstructured or semi-structured from which extraction of useful information is a difficult task. Sophisticated language processing techniques can be applied to extract symptoms and medication information from the medical documents. Clustering text documents is of high importance in the era of information explosion as data on the internet is increasing dramatically every single day. Large parts of the data are in text format and in most cases exist with no labels. Annotating the text documents manually is usually a tedious human task; although automatic annotation techniques exist, still they are not accurate. For this reason, clustering is considered as an important data mining technique in categorizing/ classifying, summarizing and organizing text documents. Document clustering techniques help in navigation and summarization of medical documents [1].

NLP-Natural Language Processing is a sub-field of Artificial Intelligence that is focused on enabling computers to understand and process human languages, to get computers closer to a human-level understanding of language [6].

- Document clustering takes the following steps:

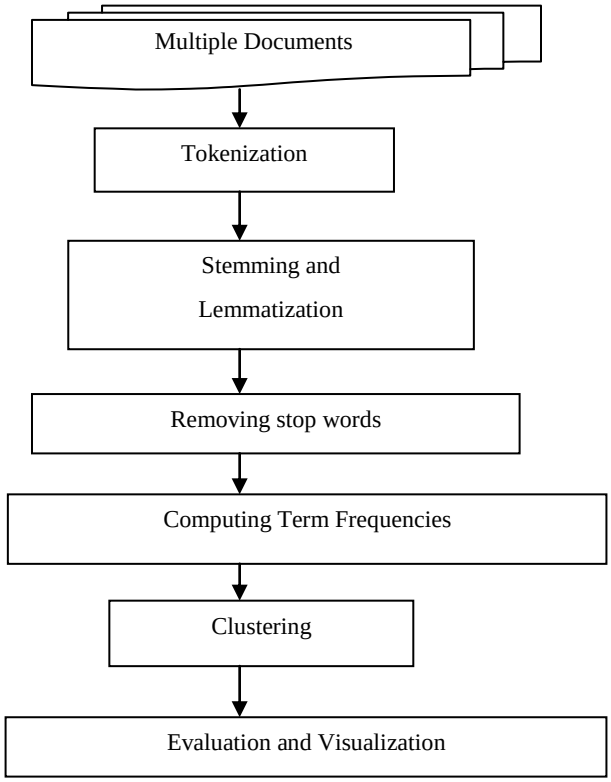

Figure 1: Steps in document clustering

- **Tokenization:**
  It is the process of dividing the text data into smaller chunks (tokens) such as words and meaningful phrases.

- **Stemming and Lemmatization:**
  Stemming is a process of stripping or dropping unnecessary characters, usually a suffix from a token.

  Ex. Healthy → Health

  Lemmatization is a process of transforming word-form into a linguistically valid base form.

  Ex. Better → Good

- **Removing stop words and punctuation**:

Some tokens are less significant than others. So it is a good practice to eliminate stop words and punctuation marks before doing further analysis.

Ex. ~~The~~ healthy kidneys make ~~a~~ hormone called erythropoietin.

- **Computing term frequencies or tf-idf**:

Feature generation can be performed after pre-processing the text data. For document clustering, one of the most common ways to generate features for a document is to calculate the

tf-idf [3] of all its tokens.

TF-IDF means term frequency-inverse document frequency. It is the weighting method which reveals the importance of a word in a document in a given corpus. The term frequency is the number of times a term occurs in a document. The term frequency for a word can be calculated as the ratio of number of times the word occurs in the document to the total number of words in the document.

The inverse document frequency helps in measuring the 'commonality' of the term across all the documents. Dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient gives the inverse document frequency.

The **tf*idf** of term t in document d is calculated as:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} * \text{idf}_t$$

- **Clustering:**

After generation of features, various clustering techniques can be applied to group different medical documents. Clustering can be defined as placing the objects into groups such that the objects in one group are more similar compared to the objects in the other group.

- **Evaluation and visualization:**

Finally, various metrics can be applied to assess the clustering models. The results can be visualized by plotting the obtained clusters.

2. ALGORITHMS

2.1 K-Means

The K-Means algorithm is a partitional clustering algorithm which aims to group a set of objects into *k* clusters [3]. *k* centroids are defined one for each cluster in such a way that it is closely related to all the objects in that cluster.

2.1.1 *Steps in K-Means Algorithm*:

1. Choose k number of clusters to be determined.
2. Randomly choose k objects as the initial cluster centers.
3. Repeat

3.1. Assign each object to their closest cluster.

3.2. Compute new clusters by calculating the mean points.

4. Until

4.1. Cluster centers do not change OR

4.2. All the objects in a cluster remain unchanged.

2.2 Agglomerative Clustering

Agglomerative clustering also called Hierarchical Agglomerative Clustering is a "bottom up" type of clustering [4]. In agglomerative clustering, each data point is defined as a cluster. Pairs of clusters are merged as the algorithm moves up in the hierarchy.

2.2.1 Steps in Agglomerative Clustering:

1. The similarity/dissimilarity information between every pair of objects is computed using distance measures.
2. Using linkage function, objects that are in close proximity are grouped together into a cluster.

2.2.2 Various linkage functions used in Agglomerative clustering:

- **Single Linkage:** In Single Linkage, similarity is calculated for the closest pair of objects [8].
- **Complete Linkage**: In Complete Linkage, similarity is calculated for the farthest away pair of objects [8].
- **Average Linkage**: In Average Linkage, similarity is calculated between groups of objects, rather than individual objects [8].
- **Ward's Method** : At each step, the process makes a new cluster that minimizes variance, measured by an index called E (also called the sum of squares index).

The linkage functions can be depicted as follows:

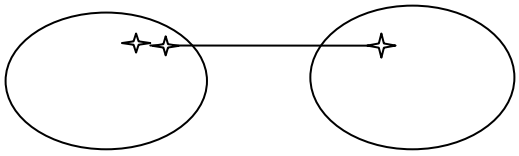

Figure 2: Single Linkage

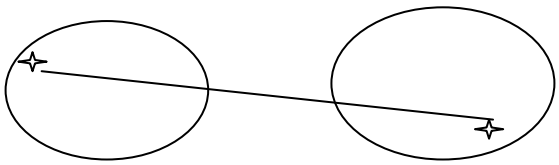

Figure 3: Complete Linkage

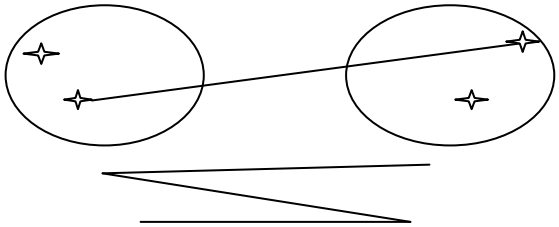

Figure 4: Average Linkage

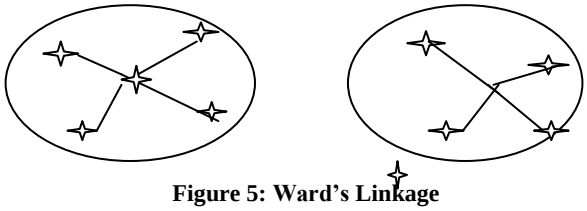

Figure 5: Ward's Linkage

2.2.3 Various Distance Measures

Different distance measures are available for calculatingthe similarity among objects. As the distance increases it resembles that the objects are more dissimilar. Some of the distance measures for the data objects x, y are as shown:

Table 1. Different distance measures

Distance Measure	Formula
Euclidean [7]	$\sqrt{\sum(x_i - y_i)^2}$
Minkowski [7]	$(\sum(\lvert xi - yi \rvert)^p)^{1/p}$
City Block	$\sum(\lvert x_i - y_i \rvert)$

x_i , y_i are the values of the i^{th} feature for x and y objects respectively. p is a constant.

2.3 Evaluation

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The silhouette value measures how an object is similar to its own cluster (cohesion) compared to other clusters (separation) [5]. The silhouette ranges from −1 to +1, where a high silhouette value indicates that the object is well matched to its original cluster. If most objects in a cluster have a high silhouette value, then it signifies that the configuration of that cluster is more appropriate.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}}$$

From the above equation it is clear that

$$-1 \leq s(i) \leq 1$$

Where, a(i) is the average distance between i and all other data within the same cluster and b(i) is the smallest average distance of i to all points in any other cluster, of which i is not a member.

3. EXPERIMENTAL RESULTS

In this paper four Document Clustering techniques are implemented and compared. They are:

- k-means
- Complete Linkage
- Wards Linkage
- Average Linkage

The same medical documents are given as input to the above four techniques with different number of clusters (k=3, 4, 5). The validation of the clusters are analyzed using silhouette value and they are tabulated in the tables Table 2 to Table 5 and depicted in the figures Figure 6 to Figure 9.

The silhouette value for kmeans algorithm with different number of clusters is as shown:

Table 2. Silhouette Value for Kmeans algorithm

No. of clusters	Silhouette Value
3	0.12449
4	0.12763
5	0.14137

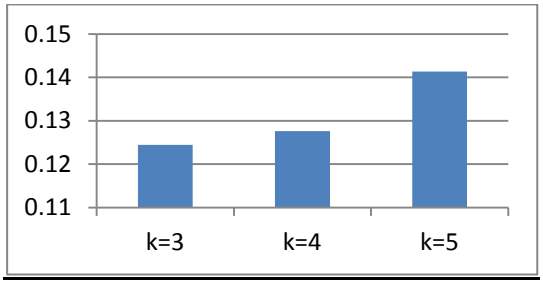

Figure 6: Silhouette value for kmeans algorithm

The silhouette value for complete linkage clustering algorithm with city block distance measure with different number of clusters is as shown:

Table 3. Silhouette Value for Complete linkage algorithm

No. of clusters	Silhouette Value
3	0.10603
4	0.11859
5	0.11859

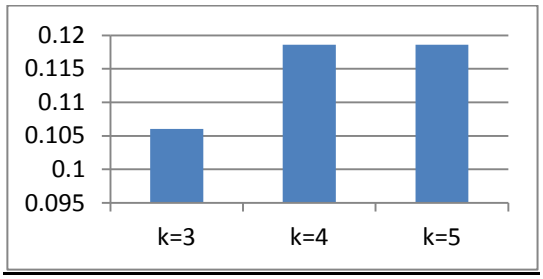

Figure 7: Silhouette value for Complete linkage algorithm

The silhouette value for wards linkage clustering algorithm with euclidean distance measure with different number of clusters is as shown:

Table 4. Silhouette Value for Wards linkage algorithm

No. of clusters	Silhouette Value
3	0.12449
4	0.12763
5	0.12763

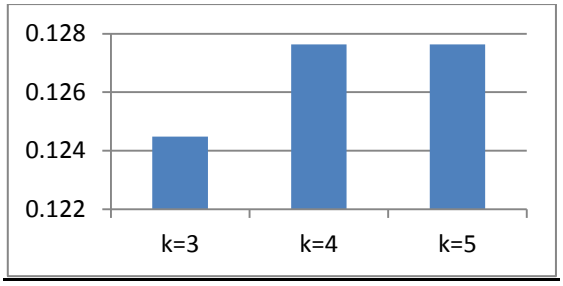

Figure 8: Silhouette value for Wards linkage algorithm

The silhouette value for average linkage clustering algorithm with minkowski distance measure with different number of clusters is as shown:

Table 5. Silhouette Value for Average linkage algorithm

No. of clusters	Silhouette Value
3	0.10443
4	0.12266
5	0.11143

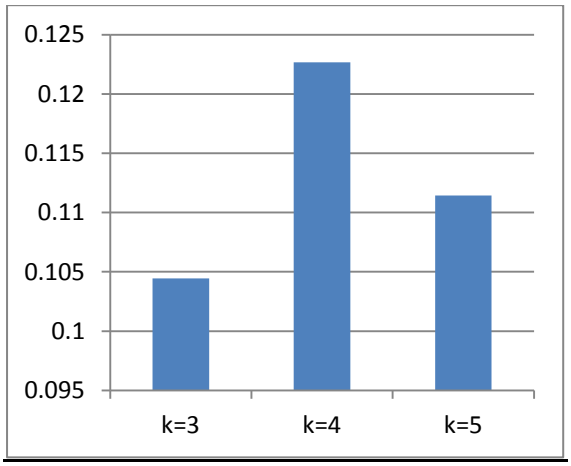

Figure 9: Silhouette value for Average linkage algorithm

3.1 Result Analysis

The silhouette value for various clustering algorithms with different number of clusters is observed and tabulated as shown:

Table 6. Silhouette value for different document clustering algorithms

S.No.	Algorithm	Silhouette-Value		
		k=3	k=4	k=5
1	k-means	0.12449	0.12763	0.14137
2	Complete Linkage	0.10603	0.11859	0.11859
3	Wards Linkage	0.12449	0.12763	0.12763
4	Average Linkage	0.10443	0.12266	0.11143

The graphical representation is as follows:

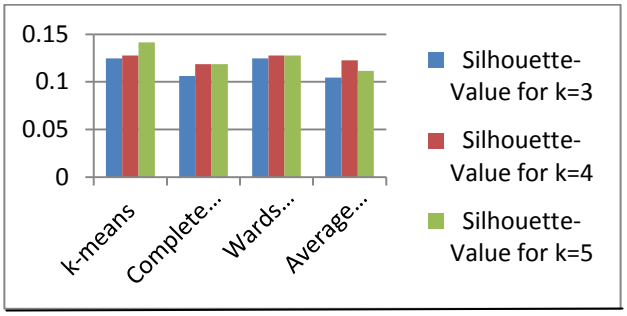

Figure 10: Analysis of different clustering algorithms based on silhouette value

From the above results it was understood that K-means and Wards Linkage algorithms perform much better than the other algorithms for all the given number of clusters in the corpus.

4. REFERENCES

[1] Yuan Ling, Xuelian Pan, Guangrong Li, and Xiaohua Hu, "Clinical Documents Clustering Based on Medication/Symptom Names Using Multi-View Nonnegative Matrix Factorization", IEEE Transactions On Nano Bioscience, Vol. 14, No. 5, July 2015

[2] K. Roberts and S. M. Harabagiu, "A flexible framework for deriving assertions from electronic medical records," Journal of the American Medical Informatics Association., Vol. 18, No. 5, pp. 568–573, 2011.

[3] T. Tarczynski, ``Document clustering-concepts, metrics and algorithms," International Journal of Electronics and Telecommunications, Vol. 57, No. 3, pp. 271_277, 2011.

[4] Aboutbl Amal Elsayed and Elsayed Mohamed Nour, "A Novel Parallel Algorithms for Clustering Documents Based on the Hierarchical Agglomerative Approach", International Journal of Computer Science & Information Technology, Vol.3, Issue 2, pp.152, Apr.2011.

[5] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, "Understanding of Internal Clustering Validation Measures", 2010 IEEE International Conference on Data Mining.

[6] Ricky K. Taira, Vijayaraghavan Bashyam, and Hooshang Kangarloo "A Field Theoretical Approach to Medical Natural Language Processing", IEEE TRANSACTIONS on Information Technology in Biomedicine, Vol. 11, No. 4, July 2007.

[7] Jasmine Irani, Nitin Pise and Madhura Phatak "Clustering Techniques and the Similarity Measures used in Clustering: A Survey", International Journal of Computer Applications (0975 – 8887) Volume 134 – No.7, January 2016.

[8] Odilia Yim and Kylee T. Ramdeen, "Hierarchical Cluster Analysis:Comparison of Three Linkage Measures and Application to Psychological Data", TQMP, Vol. 11, No. 1, 2015, DOI: 10.20982 / tqmp 11.1.p008.