International Conference on Computer Applications 2016 [ICCA 2016]

ISBN	978-81-929866-5-4		**VOL**	05
Website	icca.co.in		**eMail**	icca@asdf.res.in
Received	14 – March– 2016		**Accepted**	02 - April – 2016
Article ID	ICCA003		**eAID**	ICCA.2016.003

Review on Document Recommender Systems Using Hierarchical Clustering Techniques

Priya Mohite[1], Pravin G Kulurkar[2]
[1,2]Vidharbha Institute of Technology, Nagpur

Abstract- *We the humans are surrounded with immense unprecedented wealth of information which are available as documents, database or other resources. The access to this information is difficult as by having the information it is not necessary that it could be searched or extracted by the activity we are using. The search engines available should be also customized to handle such queries, sometime the search engines are also not aware of the information they have within the system. The method known as keyword extraction and clustering is introduced which answers this shortcoming by spontaneously recommending documents that are related to users' current activities. When the communication takes place the important text can be extracted from the conversation and the words extracted are grouped and then are matched with the parts in the document. This method uses Natural Language Processing for extracting of keywords and making the subgroup that is a meaningful statement from the group, another method used is the Hierarchical Clustering for creating clusters form the keywords, here the similarity of two keywords is measured using the Euclidean distance. This paper reviews the various methods for the system.*

Keywords: *Natural Language Processing (NLP), Hierarchical Clustering, Euclidean Distance*

1. INTRODUCTION

The keyword extraction method which is used to extract keyword from the conversation is proposed with the goal of using the keyword to retrieve, for each short conversation fragment a small number of relevant documents automatically searched and recommended to the participants. This method spontaneously recommends the documents that are related to the activity that the user is doing. Here the main focus of activity would be the conversation. Since in conversation there are multiple potential words which can relate to multiple topics. When users participate in a meeting, their information needs can be modeled as keywords that can be extracted from text based conversation and documents. These keywords then organized into subgroups and can be matched to recommend relevant document to the user.

Keyword extraction method uses Natural Language Processing. NLP is the field concerned with human (language) and computer interaction. The subsets of the keywords are obtained by using hierarchical clustering. Hierarchical clustering algorithms are either top-down (Divisive) or bottom-up (Agglomerative). The recommendation lists were prepared by ranking the documents and measuring the similarity based on the Euclidean distance of the corresponding keywords extracted from conversation fragment and documents.

Document Recommender is a system that could provide relevant documents for an ongoing discussion, intended for use in meetings. The system can be use to make business to business communication easy and more productive.

In this system the main purpose of the meeting is to felicitate direct communication between participants and here the document plays

Cite this article as: Priya Mohite, Pravin G Kulurkar. "Review on Document Recommender Systems Using Hierarchical Clustering Techniques". *International Conference on Computer Applications 2016*: 10-14. Print.

a very important role. These documents contain facts that are currently discussed but can be not exactly same. However when using the chat the user do not have time to perform such searches therefore a system that could provide relevant documents for an ongoing discussion would be helpful.

2. Review of Techniques

2.1 Recommendation Classes

There are multiple recommendation classes available which are feature, knowledge, behavior, citation, context, and ruse based. Since all the approaches are correct within their specific domain then we cannot say that a particular system is good. The technique can be chosen from user needs and computation requirements. The recommendation techniques are

2.1.1 Stereotyping

It is one of the earliest methods it was first introduced in the recommender system Grundy proposed by Rich to recommend novels to its users. The recommendation was inspired by the stereotypes from psychology that allow doctors to judge the patient on just few characteristics. This system proposed "Facets" which are the collection of the characteristics. For example to suggest novels to a male the facet contains that males prefer suspense, action thrill etc and on that bases the system generates the recommendation of the novels for them.

One major problem with stereotypes is that they can pigeonhole users. While many men may have a negative interest in romance, this is not true for all men. In addition, building stereotypes is often labor intensive, since the items typically need to be manually classified for each facet. This limits the number of items, for example, books that can reasonably be personalized

2.1.2 Content Based Filtering

The Content based filtering (CBF) is the most used and researched algorithm for the recommendation system. This technique can be compared with the star schema here there is a central node which is termed as the user modelling processes. It is filled with the interest of the user which are derived from the items. Items are nothing but the set of transactions made by the user. The items could be the email transaction, search query etc. The features of items can be the typically word-based, i.e. single words, phrases, or n-grams. Some recommender systems also use non-textual features, such as writing style, layout information and XML tags. Here only the most descriptive features are used to model an item and the features and users are given a specific weight/priority. Once these features are identifies they are stored in a vector form that contains features and their respective weights. To generate the recommendation the user model and recommendation candidates are compared. for example using the vector space model and the cosine similarity coefficient.

This technique has a number of advantages compared to the previous stereotypes they are CBF allows a user-based personalization so the recommender system can determine the best recommendations for each user individually, rather than being limited by stereotypes. CBF also requires less up-front classification work since user models are created automatically. Whereas there are some disadvantages also like it requires more computing power than stereotyping and each item is analyzed before building the user model which takes a lot of time. It also ignores the popularity and quality of the items leading to low serendipity and overspecialization.

2.1.3 Collaborative Filtering

This technique was proposed by the Goldberg et al his concept states that "information filtering can be more effective when humans are involved in the filtering process". The actual theory says that users like what like-minded users like image two users were considered like-minded if they rated items alike. When like-minded users were identified items that one user rated positively were recommended to the other userand vice versa. Compared to CBF the CF offers three advantages which are as follows. The first advantage is that CF is content independentthat is no error-prone item processing is required by the system. Secondly because humans do the ratings for the items they like and the CF takes into account real quality assessments to ensure the accuracy. Finally CF is supposed to provide serendipitous recommendations because recommendations are not based on item similarity but on user similarity i.e. the ratings.

A very general problem of CF is that here the rating matter a lot if we have to recommend a movie to any of the user here the number of users are more than the movie and the likings match the movie can be good recommendation for that user but that's not the case in documents. There can be thousands of documents and very few users so it will be difficult to optimize the users. The another disadvantage of the system is that its computing time is much more and is less scalable. In general it is termed as black box it recommend things only because some users like it. In CF manipulation is also not possible since most rated would be most recommended too.

2.1.4 Co-Occurrence

Co-Occurrence is another documentation recommendation technique here only those items are recommended that have co-occur with some source item. It was first implemented in an application called as small, here two papers that are co cited are recommended. Same was also implemented in the Amazon. In Amazon if a user buys a particular item he would also co cited item like if a user buys a cell phone he will buy it cover so the system will recommend covers to the user.

One major advantage of this system is that the main focus here is on the relatedness and not on similarity hence we can say that this technique can provide more serendipitous recommendations. Here one addition is that no access to content is needed and complexity is rather low. It is also rather easy to generate anonymous recommendations, and hence to assure users' privacy. Whereas the major disadvantage is that if the source does not occur its related recommendation will also not be shown.

2.1.5 Global Relevance

It is the simplest technique from all the systems. It is based on the one-fits-all approach and it recommends items that is having the highest global relevance. The global relevance is not constrained that is it is not user specific like we studied above the user model and the rating model. Here Global measures are used like the overall popularity. As we seen in the rating the movie is recommended on the basis of the most tickets sold and average rating but here the first choice would be user likeness then the global measures like the rating and all would be used.

It is not used as standalone system but it used as an addition to the system for ranking factor. It can be be easily joined with the CBF. There is a lot of scope for research in this technique.

2.1.6 Hybrid

As the name suggests it is the mixture of multiple techniques to get more accurate outcome. Many of the above reviewed techniques have hybrid characteristic like for example several of the CBF approaches use global relevance attributes to rank the candidates or the graphs are used to extend or restrict the recommendation system results according to the user needs. This type of hybrid techniques are called as "feature augmentation".

Since we can mix any two techniques together there is a lot of scope for the system to improve the above mixture is just the basic system for the hybrid. The very first successful implementation of this system was in TechLens it is a document recommendation system created by GroupLens. There were multiple updates of the system and are still going the first successful algorithm was given by Robin Burke which consists of three CBF variations, Two CF variations and Five Hybris approaches together.

2.1.7 Natural Language Processing

The Natural Language processing or the NLP is one of the advanced method to use in the document recommendation system. In here an intelligent system is created which is capable of keyword extraction and searching the content. The NLP has the following steps. Firstly the chat has to be analysed like for example is a meeting is going and the chat contains the topics of discussion, name of attendees etc. So the NLP system takes this system as the input and the analysis takes Place.

The second step is the Keyword Extraction, Here the keywords from the chat document is to be selected. In the language processing there are Main words and some supporting words etc and the sentence is made of keywords etc. Here the system will remove the words like is the etc and only the Nouns and The verbs are mostly stored for the keywords. For example the sentence says "This is an NLP paper". Here is and an would be removed and This, Paper, NLP is stored. Always the keywords should have some meaning else it would be difficult by the system to analysis.

The third step is keyword association here image the user has a document of nearly thousand words and the NLP processed the document and the document is cut short to only few hundreds and from the remaining the main keywords are chosen and if the matches with the Extracted keyword the document is suggested.

For Example let us assume that there are two people who are chatting. The main agenda off the meeting is to buy a land in some locality. The user A has documents for land in the same locality soo while the chat is ongoing the NLP system should take input from the chat and should extraction the keywords and after extraction it should understand the meaning of the chat that is taking place by the keywords. Once the system understands it should search for the relevant documents and should give a suggestion of that document to the user while the chat is on so that the document can be used for the meeting.

2.2 Clustering Methods

In this section we will be reviewing some of the well known clustering methods which can help us in the formation of the logic or the intelligent statement from the set of keywords. It also helps in clustering same documents together. The clustering Methods are reviewed as follows:

2.2.1 Hierarchical Methods

In this method the cluster are constructed by recursively partitioning the instances in a top down or bottom up fashion. This method can be subdivided as follows:

Agglomerative hierarchical clustering: Here each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained.

Divisive hierarchical clustering: Here all objects initially belong to one cluster. Then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters. This process continues until the desired cluster structure is obtained.

Here in this method the result is stored in a dendrogram which represents the grouping of objects and the similarity of the grouping. A cluster of data is obtained by cutting the dendrogram at desired similarity level. The division is done on some similarity measures which are as follows:

Single-Link Clustering: It is also called as the nearest neighbor or the minimum method. Here the main criteria is the distance. If a distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, the similarity between a pair of clusters is considered to be equal to the greatest similarity from any member of one cluster to any member of the other cluster.

Complete-Link Cluster: It is also called as the diameter or the maximum method. This method consider the distance between two clusters should be the longest distance from any member of one cluster to the second cluster.

Average-Link Cluster: It is also called the minimum variance method. In this method the distance between two clusters should be equal to the Average of the distance same should be with the every element from the set of second cluster element.

The main advantages of this technique is the versatility it maintain good performance on data sets containing non-isotropic clusters and the second advantage is Multiple Partition where the hierarchical methods produce not one partition, but multiple nested partitions. There are few disadvantages of the system too which are as follows. Firstly this method is unable to scale well and Hierarchical methods can never undo what was done previously. Namely there is no back-tracking capability

2.2.2 Partitioning Methods

In the partitioning method the instances are relocated by moving them from one cluster to another. It starts with the initial partitioning and continuing it. Such methods needs the number of cluster should be pre set by the user. To produce global optimization all the possible partitions should be made but since it not possible to do some of the greedy heurictics are used. Following are the various partitioning methods:

Error Minimization Algorithm: The most intuitive and most frequently used algorithms are the one which tend to work well with the isolated and compact clusters. The basic idea was to find a cluster structure that can minimize the error criteria here it measures the distance of each instance to its representative values. SSE may be globally optimized by exhaustively enumerating all partitions, which is very time-consuming, or by giving an approximate solution (not necessarily leading to a global minimum) using heuristics

Graph Theoretic Clustering: This method produce clusters via the graphs. The edge of the graph connect the instances are represented as nodes. A well known algorithm which is based Minimal Spanning tree is MST. Here the inconsistent edges are those edges whose weight is significantly larger than the average of nearby edge lengths.

2.2.3 Density Based

In the density based method the points that belongs to each cluster are drawn from a specific probability distribution whereas the overall distribution of the data is the mixture of several distributions. Here the main aim is to find the clusters and their distribution parameters. These methods are designed for discovering clusters of arbitrary shape which are not necessarily convex

$$x_i, x_j \in C_k$$

This does not necessarily imply that:

$$\alpha \cdot x_i + (1 - \alpha) \cdot x_j \in C_k$$

Here the cluster will grow till the density exceeds some threshold. That is the radius of the cluster should contain some minimal amount of nodes, here when each cluster is characterized by the local mode or the maxima of the density function these methods are called mode-seeking. In this method the majority of workload is based on the assumption that the densities of component are multivariate Gaussian or multinominal. An acceptable solution in this case is to use the maximum likelihood principle. According to this principle, one should choose the clustering structureand parameters such that the probability of the data being generated by such clustering structure and parameters is maximized. Density-based clustering may also employ nonparametric methods, such as searching for bins with large counts in a multidimensional histogram of the input instance space.

2.2.4 Model Based

In this method the main aim is to optimize he fit between the given data and some mathematical models. It is not similar to the conventional clustering techniques It find characteristic descriptions for each group, where each group represents a concept or class. The most popular model based techniques used are the decision tree and the neural networks.

Decision Tree: Here the data is represented in the form of hierarchical tree where the leaf signifies a concept and contains a probabilistic description of that concept. Several algorithms produce classification trees for representing the unlabelled data. The most well-known algorithm is COBWEB.

Neural Networks: Here the clusters are termed as neurons or the prototype also the input data is also represented by neurons which are connected too their prototyped neurons. Each connection has some weight which is better understood in the learning level of the system. A very popular neural algorithm is the Self-Organizing map (SOM).

2.2.5 Fuzzy Clustering

In the traditional Clustering approaches partition takes place in a partition each instance belong to one and only one cluster therefore the clusters in a hard clustering are disjointed. Fuzzy Clustering extends it to match the soft clustering. Now each pattern is associated with every cluster using some sort of membership function, namely, each cluster is a fuzzy set of all the patterns. Larger membership values indicate higher confidence in the assignment of the pattern to the cluster. A hard clustering can be obtained from a fuzzy partition by using a threshold of the membership value. The most popular fuzzy clustering algorithm is the fuzzy c-means or the FCM.

3. Conclusion

From all the methods we reviewed above in the paper for this project we will use the NLP technique as it is the most flexible and is less time consuming. It can also provides results much faster than any other algorithm. Whereas we will use the Hierarchical Clustering method as it easy to implement, fast, Versatile and we can have as many partitions as we want in the systems.

4. Reference

1. “Clustering Methods” by LiorRokach and OdedMaimon.
2. “Keyword Extraction from Meeting Documents for Search and Retrieval ” by Caslon Chua, Clinton Woodward
3. “NLP-based Course Clustering and Recommendation” by Kentaro Suzuki, Hyunwoo Park
4. “Efficient Bayesian Hierarchical User Modeling for Recommendation Systems” by Yi Zhang , Jonathan Koren
5. “Recommender Systems” byPrem Melville and VikasSindhwani
6. “Recommendation Systems”
7. “Content-based Recommendation Systems” by Michael J. Pazzaniand Daniel Billsus
8. “Content-basedRecommender Systems: State of the Art and Trends” by Pasquale Lops, Marco de Gemmis and Giovanni Semeraro
9. “BringingOrdertoLegalDocuments AnIssue-basedRecommendation System via Cluster Association” by Qiang Lu and Jack G. Conrad
10. “Research-Paper Recommender Systems: A Literature Survey ” byJoeranBeel, Bela Gipp, Stefan Langer, and Corinna Breitinger
11. “Automatic Tag Recommendation Algorithms for Social Recommender Systems” by YANG SONG, LU ZHANG
12. “A Document Recommendation System Blending Retrieval and Categorization Technologies” by Khalid Al-Kofahi, Peter Jackson, Mike Dahn*, Charles Elberti, William Keenan, John Duprey

Cite this article as: Priya Mohite, Pravin G Kulurkar. “Review on Document Recommender Systems Using Hierarchical Clustering Techniques”. *International Conference on Computer Applications 2016*: 10-14. Print.