Web Mining: Opinion and Feedback Analysis for Educational Institutions

Jai Prakash Verma
Assistant Professor
Intitute of Technology,
Nirma University, Ahmedabad

Bankim Patel, Ph.D
Professor & Director
SRIMCA,
UKA Trasadia University, Surat

Atul Patel, Ph.D
Professor & Dean
CMPICA,
CHARUSAT, University

ABSTRACT
Big amount of data available in the forms of reviews, opinions, feedbacks, remarks, comments, observations, clarifications, and explanations that require a robust mechanism to store, retrieve, analyze, and management. In this paper, we are proposing system that provides review or summary of above mention text data available on web for an educational institute. Due to big data size above system can be extended as real time recommendation system for Big Data analysis as future enhancement.

Keywords
Web Mining, Opinion Analysis, Feedback Summarization, Sentiment analysis, k-mean clustering, Recommendation System etc.

1. INTRODUCTION
The web is becoming the important part of people's regular life. It can be viewed as the largest database available and presents a challenging task for effective storage and retrieval. Here the term database used quite loosely because there is no authentic structure or schema to the web data [4, 6, 7]. Web Mining is mining of data associated to the World Wide Web. This may be the data actually present in Web pages or data associated to web activity. Web Data Mining can be classified basically in three classes: Web content mining, Web Usage Mining, Web structure mining [6]. Opinion mining and sentiment analysis is covered in the area of web content mining [2, 3, 4, 7]. Positive and negative opinions represent the quality, reliability, durability, and feasibility about the product or website. Knowledge extracted from these opinions can influence directly on users decision making process about the particular product, their category, and way of transaction facility provided by website. Irrelevant opinions are the opinions that are not relevant to the product or website. Some time viewers of website enter some irrelevant remarks or texts in the available fields on website generates these types of irrelevant opinions. False and spam opinion are generated for promoting, some time negative promotion about a product or category of products based on the business interest of different stakeholders. These types of opinions are written or produced by notorious uses as well as sponsored professional users to promote or demote a particular product or website [2, 4, 7]. In this paper we are discussing the issue of opinion mining and summarization of feedback, reviews, and remarks given by different stakeholders for the functioning of a university and educational institution. In this paper we are proposing an intelligent and predictive model that prepare summary for all the reviews, feedback and remarks given by students.

The paper is worded as follows: Section 2 presents proposed system and its different components. In Section 3, the important issues in the area of opinion mining. Section 4 presents methodology used to implement the proposed system. Section 5 presents result analysis. Section 6 concludes the work and presents future work. Due to big data size (Big Data) above model can be implemented on distributed system using the concept of Hadoop and Map-Reduce as future enhancement in this work.

2. PROPOSED WORK
In our universities and educational institutions big amount of data available in the forms of reviews, opinions, feedback, remarks, comments, observations, clarifications, and explanations that require a robust mechanism to store, retrieve, analyze, and management. In our University system these types of feedbacks on syllabus, facilities, teaching learning etc. are given by different stakeholders like students, parents, industry experts, research experts, and visitors. One remark field is also provided where they can write anything related to functionalities of the system. Problem is how to summaries these remarks. Currently this process is doing manually. Figure 1 represent proposed model for analyzing different types of feedbacks, reviews, and remarks. In this model data extracted from different data sources like university web site, several online feedback systems running in university, and feedback and suggestions box that are digitalize by staff. These collected data that are in the form of set of review, feedback, remarks can be cleaned in data preprocessing step. Missing and noisy information can be removed because many reviews and feedback are given without the truthfully by the different stakeholders. In the Ontology Learning step we are identifying unique and effective terms that make significance in review and feedback for decision making or action taken. After it we apply grouping for similar or synonym terms. In the tokenization step we assign weight to each term that will be used in next step to calculate weight for each review or feedback. Based on token assign to each term we select frequently used term for each review as per threshold value decided by domain expert in Feature Selection step. Now we can apply data mining technique for cluster analysis and summary analysis that shows the summary and important feedbacks form available huge amount of reviews, feedback, and remarks for decision making and action taken.

3. RELATED WORK
Opinion mining analyzes people's remarks, appraisals, opinions, and emotions based on different data mining techniques. These opinions and remarks are key influencers of visitors and customers of websites. Because of the opinion the user can take decision about the web site facility as well as product that he/ she will be going to purchase [6].Opinions are categories as Positive Opinion, Negative Opinion, Irrelevant Opinion, False and Spam Opinion. Following are related work that we identify in the area of web data mining for opinion or sentiment analysis.

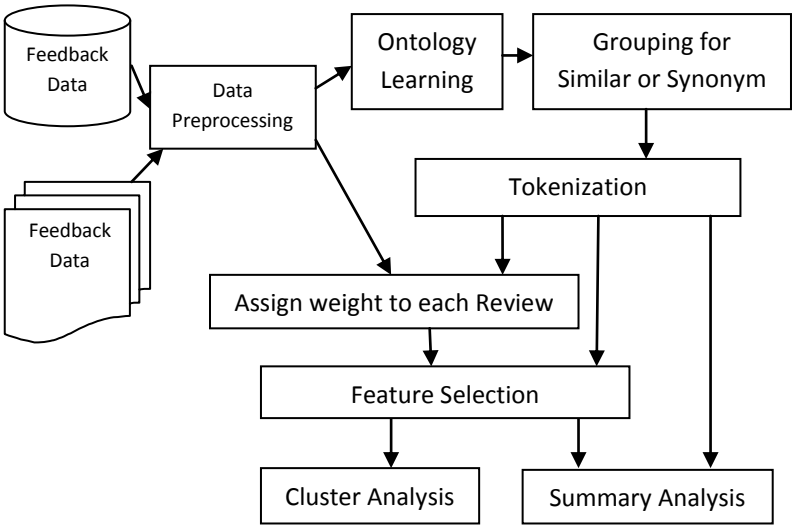


Figure 1: Consecutive steps for proposed student's feedback analysis model

Tushar Ghorpade, Lata Ragha, [1] proposed a system to improve training dataset by adding weight to the world extracted from reviews and opinions using JAPE mathematical technique of NLP (Natural Language Processing) concept. It also proposes a machine learning algorithm to classify the positive or negative reviews using NLP and Bayesian classification.

Bishas Kaur, Aarpit Saxena and Sanjay Singh [2], proposed a predictive model to analysis the social site's or blog's reviews and opinions using DBSCAN and SDC clustering techniques. It also shows analysis results of comparing three different types of review collected from a social site.

Nitin Jindal and Bing Liu [3] analyze the spam reviews posted by different types of users intensely. The most important in spam is web spam that can be categorized in two type web content spam and web link spam. Web spam can increases web page ratings that can influence the web search to visit the webpage.

Nitin Jindal [4] deals with a restricted problem, i.e., identifying unusual review patterns which can represent suspicious behaviors of reviewers in the paper. They formulate the problem as finding unexpected rules. The technique is domain independent. Using the technique, they analyzed an Amazon.com review dataset and found many unexpected rules and rule groups which indicate spam activities.

Arjun Mukherjee [5] focuses on this task and proposes an effective technique to detect spam reviewer groups. This paper proposed an effective technique to detect spammer groups who work together to write fake reviews. And user-agreement study showed that the technique is promising.

Zhongwu Zhai [8] focuses on summarization of reviews based on opinion features and aspects using semi-supervised learning approach. To summarize a document paper proposes to make different domains of words that are synonyms.

4. METHODOLOGY USED

Here we are proposing content based clustering technique to analysis of feedback and remarks given by students. To perform content clustering technique following steps are followed.

Step1. Data Preprocessing: In data preprocessing step we clean the data provided by university student feedback system. Some time students set a blank review or review with very small text. For this experiment we removed all the missing, very small sized reviews.

Step2.Ontology Learning: In this step our model selects features based on the unique words that are making effect in the review or feedback. The ontology learning framework [1, 10, 11] can help to build an effective feature vector. Ontology learning framework consists of several steps like term extraction, ontology building, and ontology pruning. Term extraction performed to identify unique terms from the document. Then using linguistics analysis overcomes the limitation of poor quality terms that are not effective in feedback or review. Ontology building derives static and procedural knowledge in the form of a hierarchy of frequent domain concepts and a hierarchy of web service functionalities. Ontology pruning filters potentially ineffective words that will be not used in review or feedback analysis [1, 10, 11]. For experiment we are finding set of unique word that are making effect in review and feedback using a program in java. Same time throwing away certain characters such as punctuation and removing stop words from the list, generated by the java program.

Step3. Grouping similar or synonym words Lexical similarity based on WordNet is commonly used in the NLP (Natural Language Processing) to determine the similarity of two words [8, 12]. It is another piece of knowledge that can be used for grouping similar or synonym words from above generated list of words. For example, "picture" and "image" has very high similarity in WordNet.

Step4. Tokenization: After collecting list of words that is effective in review or feedback analysis has to be tokenize. TF and IDF calculation is used to select a certain numbers of top ranked terms based on term frequency – inverse document frequency (tf - idf) computation to form a document vector for each review. TF – IDF weighting combine the definitions of term frequency and inverse document frequency to produce a composite weight for each term in each review and feedback. The (tf-U-idf) weighting scheme is given by (tf –idftd = tftd x idft) which assign to term t a weight in document d [2]. For experiment we assign weight to each term in the review depends on the number of occurrences of the term in the document.

Step5. Assign weight for each feedback or remark: For assigning weight to each review we are using a java program. Based on the above word list and their weights program we assign the series of weights to each review related to their respective terms. Figure 2 shows graph between all the reviews and their member word's token assigned in tokenization step. Also we calculate cumulative weight to each review by adding weight to their member terms. Figure 3 shows graph between all reviews and their weight-age, number of words in a review or feedback.

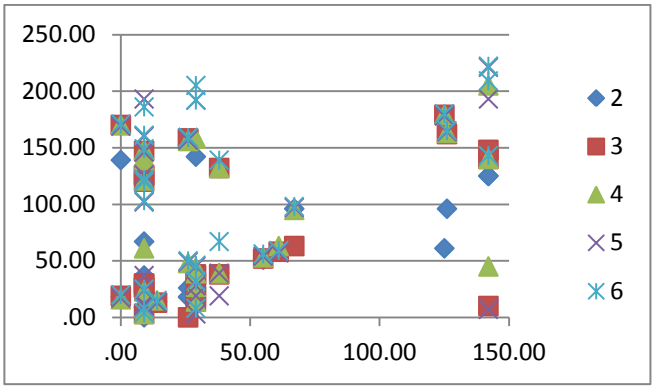


Figure 2: Graph between all reviews and their word tokens

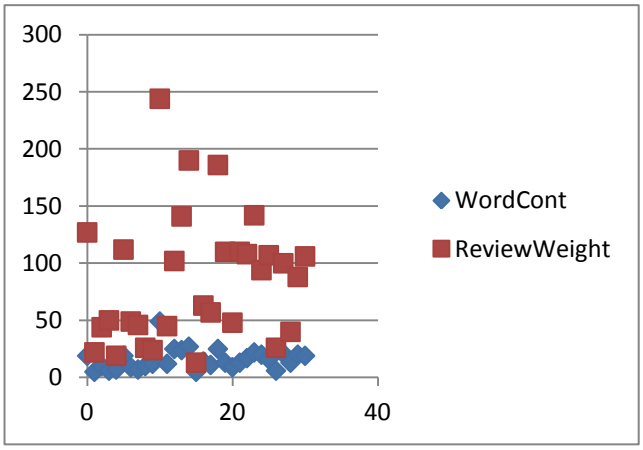


Figure 3: Graph shows all sentences with their weight-age.

Step6. Feature Selection In this step we are finding features for each review for applying cluster analysis and summary analysis. First we arrange the entire terms in each review in ascending order of their weights. Then calculate mode for set of numbers that represents total terms in each review for finding threshold value. This value can be used to choose number of feature for analysis. Now we select number of terms for each review that is equal to mode calculated previously. Here selected terms represent features for each review that can be used for cluster and summary analysis.

5. RESULT ANALYSIS

Step7. Cluster analysis using K Mean, DBSCAN: Clustering is similar to classification in that data are grouped. However, unlike classification, the subsets are not predefined. Instead, the grouping is accomplished by finding similarities between data according to characteristics found in the actual data. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other cluster. Clustering algorithms can be categories in different methods like Partitioning method, Hierarchical method, Density-based method, Grid- based method. Here we are analyzing dataset prepared in previous steps using k-mean (partitioning method), DBSCAN (density-based method) algorithm [13, 14].

K- Mean Cluster Analysis: K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached. As such, it may be viewed as a type of squared error algorithm, although the convergence criteria need not be defined based on the squared error. A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among elements in different clusters is achieved simultaneously [13, 14]. The cluster mean of cluster Ki = {ti1, ti2, ti3, ti4…} is defined as:

$$m_i(cluster\_mean) = \frac{1}{m}\sum_{j=1}^{m} t_{ij}$$

Similarity Functions: Here we are presenting brief overview of different similarity function that can be used for finding similarity between different reviews and feedback in k-mean clustering.

Euclidean Distance Similarity: Euclidean distance is broadly used in clustering problems, including clustering text. It is also default distance measure used with the K-means algorithm. Measuring distance between text documents, given two documents da and db represented by their term vectors ta and tb respectively, the Euclidean distance of the two documents is defined as

$$D_E(\vec{t_a},\vec{t_b}) = (\sum_{t=1}^{m} | w_{t,a} - w_{t,b} |^2)^{1/2}$$

Where the term set is T ={t1,t2 ….,tm}. as mentioned previously in step 4, we use the tfidf value as term weights, that is wt,a = tfidf(da,t) [16].

Cosine Similarity: When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Cosine similarity is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval applications and clustering too. Given two documents $\vec{t_a}$ and $\vec{t_b}$, their cosine similarity is

$$SIM_C(\vec{t_a},\vec{t_b}) = \frac{\vec{t_a}.\vec{t_b}}{|\vec{t_a}| \times |\vec{t_b}|}$$

Where $\vec{t_a}$ and $\vec{t_b}$ are m-dimensional vectors over the term set T = {t1, t2 …., tm} [16].

Jaccard Coefficient Similarity: For text documents, the Jaccard Coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms. The formal definition is:

$$SIM_j(\vec{t_a},\vec{t_b}) = \frac{\vec{t_a}.\vec{t_b}}{|\vec{t_a}|^2 + |\vec{t_b}|^2 - \vec{t_a}.\vec{t_b}}$$

The Jaccard coefficient is a similarity measure and ranges between 0 and 1 [16].

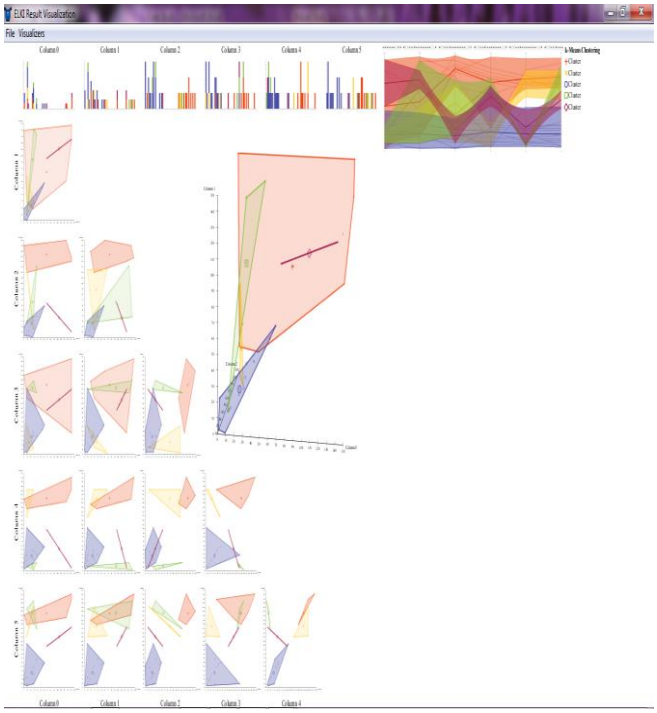
Result from SPMF: A Sequential Pattern Mining Framework (SPMF) is an open-source data mining platform written

in Java. It is distributed under the GPL v3 license. It offers implementations of 52 data mining algorithms for: sequential pattern mining, association rule mining, frequent itemset mining, sequential rule mining, clustering. It can be used as a standalone program with a simple user interface or from the command line. K-Means is one of the most famous clustering algorithms. It is used to separate a set of vectors of double values into groups of vectors (clusters) according to their similarity. In this implementation the Euclidian distance is used to compute the similarity. K-Means takes as input a set of vectors containing one or more double values and a parameter K (a positive integer >=1) indicating the number of clusters to be created [9]. For experiment here we are taking k=5 for identifying 5 reviews that can represent all sets of review and feedbacks, following figure 4 show the clusters generated by SPMF tool using k-means clustering.

Cluster – 1:
[126.0,96.0,162.0,163.0,164.0,165.0][0.0,139.0,170.0,170.0,169.0,171.0][142.0,147.0,148.0,205.0,221.0,222.0][125.0,61.0,179.0,179.0,178.0,180.0]
Cluster – 2:
[9.0,0.0,3.0,8.0,10.0,4.0][14.0,14.0,13.0,15.0,13.0,15.0][0.0,18.0,19.0,16.0,17.0,20.0][9.0,9.0,26.0,23.0,24.0,25.0][29.0,30.0,31.0,27.0,28.0,32.0][29.0,37.0,38.0,39.0,19.0,8.0][29.0,26.0,18.0,44.0,45.0,46.0][26.0,26.0,0.0,48.0,49.0,50.0][55.0,51.0,52.0,53.0,54.0,56.0][61.0,58.0,58.0,63.0,57.0,59.0][38.0,39.0,38.0,39.0,39.0,67.0][9.0,9.0,30.0,4.0,102.0,103.0][9.0,9.0,29.0,142.0,37.0,8.0]
Cluster - 3.
[142.0,125.0,10.0,140.0,7.0,143.0][67.0,96.0,63.0,95.0,97.0,98.0][9.0,9.0,26.0,139.0,12.0,186.0][29.0,142.0,37.0,158.0,3.0,205.0][38.0,132.0,132.0,132.0,19.0,139.0]
Cluster - 4. [142.0,125.0,140.0,45.0,193.0,209.0]
Cluster - 5.
[9.0,37.0,147.0,148.0,146.0,149.0][26.0,18.0,158.0,156.0,157.0,159.0][9.0,29.0,120.0,121.0,160.0,161.0][26.0,18.0,158.0,156.0,157.0,159.0][9.0,9.0,125.0,61.0,126.0,120.0][9.0,67.0,126.0,3.0,193.0,122.0][29.0,29.0,14.0,14.0,192.0,192.0]
Figure 4: cluster for k=5 using SPMF: A Sequential Pattern Mining Framework

Result from ELKI (Environment for Developing KDD-Application Supported by Index- Structures): In ELKI, data mining algorithms and data management tasks are separated and allow for an independent evaluation. This separation makes ELIKI unique among data mining frameworks like WEKA or YALE and framework for index structure like GiST. At the same time, ELIKI is open to arbitrary data types, distance or similarity measure, or file formats. The fundamental approach is the independence of file parsers or database connections, data types, distances, distance functions, and data mining algorithms [15]. Here we are selecting clustering.kmeans.KMeansLloyd algorithm with number of cluster k=5. Following figures 5(a,b,c) are showing results from elki tool using K-mean clustering with above mansion parameters.

Figure 5 (a): K-Mean clustering using Elki (selecting clustering.kmeans.KMeansLloyd algorithm with number of cluster k=5) .



Cluster Analysis using DBSCAN (Density – Based Spatial Clustering of Application with Noise):

The density of an object ‘O’ can be measured by the total number of objects closed to ‘O’. DBSCAN finds core objects, that is, objects that have dense neighborhoods. It connects core objects and their neighborhoods to form dense regions as clusters. A user specified parameter $\epsilon > 0$ is used to specify the radius of a neighborhood we consider for every object. The ϵ-neighborhood of an object ‘O’ is the space within a radius ϵ centered at ‘O’. To determine whether a neighborhood is dense or not, DBSCAN uses another user-specified parameter, MinPts, which specifies the density threshold of dense regions. An object is a core object if the ϵ-neighborhood of the object contains at least MinPts objects.

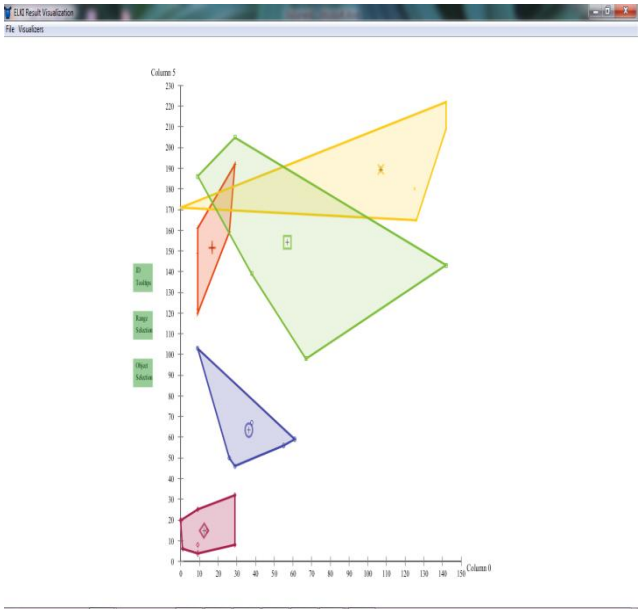


Figure 5 (b): K-Mean clustering using Elki

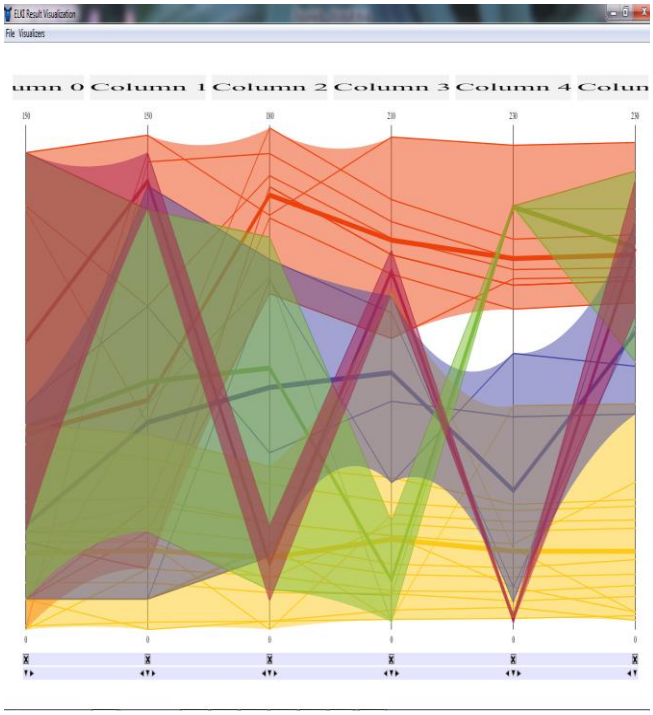


Figure 5 (c): K-Mean clustering using Elki

Core objects are the pillars of dense regions [13]. We are using Elki for analyzing DBSCAN clustering with the dbscan.epsilon = 10, and dbscan.minpts = 4. Figure 6 shows the results of DBSCAN clustering on the dataset with attributes: no of terms in each review and their weight.

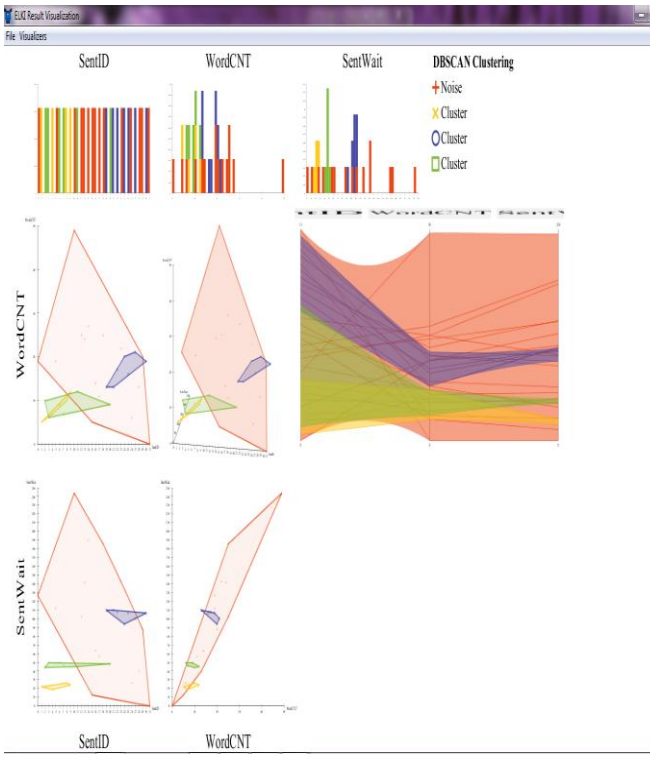


Figure 6: DBSCAN clustering using Elki

Step8. Summary Analysis:

In step 5 we calculate weight of terms for each review. Here we find set of unique terms that have highest weight in their reviews. Sense of the selected terms is representing summary of all feedback and reviews that can be used for decision making or action taken.

6. CONCLUSION AND FUTURE WORK

In this paper have proposed opinion and feedback analysis system that can help educational institutions to summaries different types of feedback and reviews. The proposed system generates the set of selected reviews as a summary of overall feedback of large data set. In this system we are using Sequential Pattern Mining Framework (SPMF) and Elki tool for clustering analysis. Results generated by these tools are shown in the different graph and found satisfactory, comparing with manually selected reviews and feedback. Due to big data size above system can be extended as real time recommendation system for Big Data analysis as future enhancement.

7. REFERENCES

[1] Tushar Ghorpade, Lata Ragha, 2012, Featured Based Sentiment Classification for Hotel Reviews using NLP and Bayesian Classification, 2012 International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India

[2] Bishas Kaur, Aarpit Saxena and Sanjay Singh, 2012, Web Opinion Mining for Social Networking Sites, CCSEIT-12 October 26-28, 2012, Coimbatore [Tamil Nadu, India]

[3] Nitin Jindal and Bing Liu, 2008, Opinion Spam and Analysis, WSDM'08, February 11-12, 2008, Palo Alto, California, USA

[4] Nitin Jindal, Bing Liu, and Ee-Peng Lim, 2010, Finding Unusual Review Patterns Using Unexpected Rules

[5] Arjun Mukherjee, Bing Liu, Junhui Wang, Natalie Glance, Nitin Jindal, 2011, Detecting Group Review Spam

[6] Bing Liu, 2011, Web Data Mining Exploring Hyperlinks, Contents, and Usages Data, Springer

[7] Bing Liu, 2012, Sentiment Analysis andOpinion Mining, Springer

[8] Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia, 2011, Clustering Product Features for Opinion Mining WSDM'11, February 9–12, 2011, Hong Kong, China. Copyright 2011 ACM

[9] Sequential Pattern Mining Framework, [Available Online], http://www.philippe-fournier-viger.com/spmf/index.php.

[10] Vladimir Oleshchuk, Asle Pedersen, "Ontology Based Semantic Similarity Comparison of Documents", Proceedings of the 14th International Workshop on Database and Expert Systems Applications (DEXA'03)" © 2003 IEEE. pp.735-738.

[11] Wu Di1, Li Xiaojing2, Zhang Chengwei3 "The Design of Ontologybase Semantic Label and Classification System of Knowledge Elements", 2011 International Conference on Uncertainty Reasoning and Knowledge Engineering, 978-1-4244-9983-0/11 ©2011 IEEE. pp. 95-98.

[12] Pedersen T, 2010, Information Content Measures of SemanticSimilarity Perform Better Without Sense-Tagged Text. In Proceedings of NAACL HLT, 2010.

[13] Jiawei Han and Micheline Kamber. Data Mining: concept and Techniques, Elseveir Pubilcation.

[14] Margaret H, Dunham, Data Mining: Introductory and Advance Topics, Pearson Education.

[15] Environment for Developing KDD- Application Supported by Index- Structures, [Available Online], http://elki.dbs.ifi.lmu.de/wiki.

[16] Anna Huang, 2008, Similarity Measures for Text Document Clustering, NZCSRSC 2008, April 2008, Christchurch, New Zealand.