Comparative Study of Various Clustering Algorithms and Sentiment Analysis Methodologies

Nishi Shah[1], Neil Parekh[2], Kshitij Jain[3]
[123]Final year Student
Shah and Anchor Kutchhi Engineering College
Contact no. 7506625719
[1]shahnishi0911@gmail.com,
[2]neilparekh123@gmail.com,
[3]jainkshitij011@gmail.com,

Bemila Theres[4]
Assistant Professor
Shah and Anchor Kutchhi Engineering College
[4]sakec.bemilat@gmail.com,

Vaibhav Vasani[5]
Assistant Professor
Shah and Anchor Kutchhi Engineering College
[5]sakec.vaibhavv@gmail.com

Abstract:- Sentiment analysis is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities and their attributes. It is also called as opinion mining. It represents a large problem space. This survey paper describes techniques and approaches that promise to directly enable opinion-oriented information seeking systems. An attempt has been made to discuss in details various approaches to perform a computational treatment of sentiments and opinions. It also discusses various approaches for sentiment analysis and clustering techniques, like K-L information regularization, Markov chain, adjective context co-clustering, maximum entropy model, MRA approach, partially supervised alignment model their strengths and drawbacks are touched upon.

*****

1. INTRODUCTION

Clustering deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is an unsupervised collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

K-L Information regularization is a method which is a fuzzy variant of multinomial mixture density estimation. Multinomial mixtures are a probabilistic model for co-clustering of co-occurrence matrices. [1] A Markov chain is a stochastic process with the Markov property. The term "Markov chain" refers to the sequence of random variables such a process moves through, with the Markov property defining serial dependence only between adjacent periods (as in a "chain") [2]. Adjective-Context-co-clustering is a method in which the object is represented by pair of aspect and its quality or value [3].

A Weighted Maximum Entropy model is a method in which weights of different words are used and its classification is based on these weights. [4]

Using open source tools introduces the basic techniques for text mining, using combination of a set of standard commands, small codes, and generic tools provided as the open-source software's [5].

In information retrieval, tf–idf, frequency–inverse document frequency, is a numerical value that is intended to reflect how important a word is to a document in a corpus. It is often used as a weighting factor in information retrieval and text mining. [6]

Partially supervised alignment model can capture opinion relations more precisely through partial supervision from partial alignment links. [7] Mutual reinforcement approach is in which customer reviews are mined quantitatively [8].

Co-clustering is a data mining technique, which allows simultaneous clustering of the rows and columns of a matrix. It is also known as bi-clustering. It can be seen as a method of co-grouping two types of entities simultaneously, based on similarity of their pair wise interactions.

Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. It is widely applied to reviews and social media for a variety of applications, which ranges from marketing to customer service

2. LITERATURE SURVEY

Katsuhiro Honda, Shunnya Oshio and Akira Notsu, had proposed a new fuzzy co-clustering model, which is a fuzzy variant of multinomial mixture density estimation. Multinomial mixture is a probabilistic model for co-clustering of co-occurrence matrices and the proposed method extends multinomial mixtures so that the degree of fuzziness can be tuned in a similar manner to K-L information-based FCM. K-L information based FCM is the fuzzy extension of the k-Means algorithm. The goal of co-clustering is to extract object-item pair wise structures from co occurrence matrices. [1]

Lazhar Labiod and Mohamed Nadif proposed a new framework which is based on iterative procedure looking up an appropriate approximation of the data matrix A by using

two stochastic similarity matrices from the set of rows and the set of columns. This process converges to a steady state where the approximated data ‘ˆA’ is composed of ‘g’ similar rows and ‘l’ similar columns. Furthermore, they showed that their approach is related to a Markov chain model. [2]

Kevin Raison, Noriko Tomuro, Steve Lytinen, and Jose P. Zagal proposed a method to represent such an object by pairs of aspect and its quality or value, for example “great graphics”. The derived co-clusters are pairs of row clusters × column clusters. By examining the derived co-clusters, they discovered the aspects and their qualities which the users care about strongly in games. A better approach is to represent each aspect and its quality together, for instance “great graphics” and “horrible graphics”. Evaluation of clustering is difficult because, unlike classification, there is no category to which the result can be measured for accuracy. Linguistic characteristics of the game domain, for example adjectives modified with nouns more often, in conjunction with the sentiment need to be checked. [3]

Kostas Fragos, Yannis Maistros, Christos Skourlas, presented a method to apply Maximum Entropy (ME) modeling for text classification by using weights for both to select the features of the model and to emphasize the importance of each one of them in the classification task. The principle underlying ME is that the estimated conditional probability should be as uniform as possible, that is, have the maximum entropy. They applied the X square test in data for feature selection and the weighting scheme, then the maximum entropy modeling and the Improved Iterative Scaling (IIS) algorithm. [4]

Jun Iio introduced the basic techniques for text mining, using combination of a set of standard commands, small code, and generic tools provided as the open-source software. They have discussed three types: Scraping Text Data from web, Splitting Word Fragments, and Visualization by Word Cloud and they have also used Hierarchical clustering. [5]

Hugo Jair Escalante, Mauricio A. García-Limón, etal proposed a genetic program which aimed at learning effective Term-Weighting Schemes (TWS) that can improve the performance of current schemes in text classification and give rise to discriminative TWSs. Genetic Programming (GP) is used as optimization strategy, where each individual corresponds to a tree-encoded TWS. The proposed genetic program explores the search space of TWSs. [6]

Kang Liu, Liheng Xu, and Jun Zhao proposed an approach based on the partially-supervised alignment model. A graph-based co-ranking algorithm is then exploited to estimate the confidence of each candidate out of which candidates with higher confidence are extracted as opinion targets. This method proposes that both opinion targets and opinion be detected. Opinion relations between opinion targets and opinion words are captured using the word alignment model and the Opinion Relation Graph is also constructed. [7]

Weiping Wang and Yuanzhuang Zhou introduced Mutual Reinforcement Approach (MRA) approach in which customer reviews are mined quantitatively. Two main traditional ways are shown to do an accurate evaluation: Expert and Questionnaire. The advantage of evaluation with this approach is that it is customer-oriented hence more credible and trustworthy. [8]

3. COMPARATIVE ANALYSIS

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature level—whether the expressed opinion in a document, a sentence or feature is positive, negative, or neutral. Sentiment analysis is done by classifying the document and then applying clustering technique. Content-based classification is classification in which the weight given to particular subjects in a document determines the class to which the document is assigned. Two different approaches studied in classification are Maximum Entropy classifier and Term Weighting Scheme. (TWS) Clustering is an unsupervised learning.

Table no 2.1Comparative studies of classification algorithms.

Method / **Parameters**	**Max Entropy**	**TWS**
Methodology	Maximum Entropy	Genetic Programming
Accuracy	Greater than 70%	Better accuracy
Reliability	Very High	Less
Applicability	Many domain	Text domain
Handling high dimensionality	No	No
Size of data set	Quite Large	Large

As observed from table 2.1 Maximum Entropy method is more reliable since it works with many different domains. The accuracy of Maximum Entropy is better than TWS since it give better results. Maximum Entropy handles large databases. TWS is used only in text domain since it is less reliable.

Table no 2.2 Comparative studies of clustering algorithms.

Method Parameters	K-L method	Unified Framework	Adjective Context
Methodology	Regularization	Visualization	Co-clustering
Accuracy	Quite Accurate	Greater than 90%	Very Accurate
Reliability	High	Very less	Medium
Applicability	Many domains	Single domain	Game domain
Handling high dimensionality	Yes	No	Yes
Size of dataset	Large	Large	Very Large

As observed from table 2.2 K-L method is reliable choice since it works with many domains. The accuracy of adjective context method is better since it provides good results as compared to K-L method. Adjective context method allows working with very large databases. Unified framework is highly accurate, when working on single domain since it cannot handle high dimensionality. Adjective context method is applicable only for game domain.

4. CONCLUSION AND FUTURE WORK

Our goal in this survey has been to cover techniques and approaches of classification and clustering that directly enable opinion-oriented information-seeking systems.

A possible future work is to adopt the deterministic annealing approach by exploiting the controllable fuzzification penalty. A more scalable version of ISMA using only matrix vector multiplication can be developed. Future work directions include studying the suitability of the proposed approach to learn weighting schemes for cross domain text classification. Considering additional type of relation between words, such as topical relations, in Opinion Relation Graph will be beneficial for co-extracting opinion targets and opinion words.

REFERENCES

[1] FCM-type Fuzzy Co-clustering by K-L Information Regularization by-Katsuhiro Honda, ShunnyaOshio and Akira Notsu.

[2] A Unified Framework for Data Visualization and Co-clustering by-Lazhar Labiod and Mohamed Nadif.

[3] Extraction of User Opinions by Adjective-Context Co-clustering for Game Review Texts by-Kevin Raison, Noriko Tomuro, Steve Lytinen, and Jose P. Zagal.

[4] A Weighted Maximum Entropy Language Model for Text Classification by-Kostas Fragos, YannisMaistros, Christos Skourlas.

[5] Basic Techniques in Text Mining using Open-source Tools by-Jun Iio.

[6] Term-weighting learning via genetic programming for text classification by-Hugo Jair Escalante, Mauricio A. García-Limón, Alicia Morales-Reyes, Mario Graff, Manuel Montes-y-Gómez, Eduardo F. Morales, José Martínez-Carranza.

[7] Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model by-Kang Liu, Liheng Xu, and Jun Zhao

[8] E-Business Websites Evaluation Based on Opinion Mining by-Weiping Wang Yuanzhuang Zhou.