Volume 4, Issue 6, June 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Competent Approach for Extracting and Visualizing Web Opinions Using Clustering

Senbagavalli M*
Assistant Professor
Jayam College of Engg&Tech.
India

Shobana K
Assistant Professor
Prince shri venkateshwara padmavathy Engg.College
India

Dr.G.Tholkappia Arasu
Principal
AVS Engineering College
India

Abstract - Huge amount of Web opinions are available in the social sites due to the development of web Communication. Web opinion acts as a boundary between the Web users and Internet. It allows the users to communicate and articulate their opinions without eye to eye contact. Nevertheless the clustering and visualizing of the web opinion is a not a trouble-free task. The obtainable Document Clustering Algorithm differs from the Web opinion clustering Algorithm. Basically, the web opinions are derived from the social networks which are user generated content. Visualizing the social network helps in preventing the Crimes, Terrorists activities and improving the Business, Political Activities in an efficient way. The scalable distance based Clustering Algorithm enables the recognition of topics within discussions in web social networks and their growth. The predefined set for clustering is valuable in web opinion clustering. In Existing Scalable Distance Based Clustering Algorithm for Web opinion Clustering has its own Limitation that is the Macro and Micro Level Accuracy. In this Paper, We improve the accuracy level by combining the SDC Algorithm with the proposed MaxEnt Re-ranking Method. When we compare MaxEnt Ranking method with p-norm Push Ranking Method, the MaxEnt Ranking method Provides Higher Accuracy and recall level.

***Keywords:** Clustering, Scalable Distance Based Clustering, opinion mining, Re-Ranking, DBSCAN.*

I. Introduction

Opinion Mining is the process to extract the opinions expressed in user-generated content. Basically the opinions are categorized into two types:

1. Direct - Sentiment Expressions on some objects.
Eg: products, topics, persons etc.

2. Comparisons to find the similarities and differences between more than one object.
Eg: car x is cheaper than car y.

Properties of web opinions (that do not exist in regular documents):

- The content of the message is fewer focused.
- Messages are usually short in length ranging from a few words to couple paragraphs.
- The terms used in the messages are thin because different users may use different conditions to discuss the same topic.
- The volume of web opinion messages is vast and ever increasing in a massive rate.

The conventional document clustering techniques that work well in clustering regular documents usually do not work well in web opinion clustering. The conventional clustering characteristics like assigning all documents into clusters or having predefined set of clusters may not be applicable to web opinion content Analysis.

A. Sentiment Analysis Applications

- Opinions are central to almost all human activities because they are key influencers of our behaviors. Whenever we need to make a decision, we want to know others' opinions.
- In the real world, businesses and organizations always want to find consumer or public opinions about their products and services.
- Individual customers also want to be familiar with the opinions of existing users of a product before purchasing it, and others' opinions about political candidates before making a voting decision in a political election.
- In the past, when an individual needed opinions, he/she asked friends and family. When an organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups.
- Acquiring public and consumer opinions has long been a huge business itself for marketing, public relations, and political campaign companies.
- With the explosive growth of social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision making.

- Nowadays, if one wants to buy a consumer product, one is no longer limited to asking one's friends and family for opinions because there are many user reviews and discussions in public forums on the Web about the product.

B. Clustering

It is the task of conveying a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters. It is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Clustering can be measured the most important unsupervised learning problem. Cluster is usually useful to reduce the load on a particular server. Basically, clustering can be defined as the use of more than one computer/ server that can work together. In this type of architecture, multiple servers are liked to one another and have the capability of handling workloads. It helps to offer continued working and offer 100% uptime. All the machines in a cluster are involved in the operations at any given point of time. The cluster of servers offer fault tolerance. It is also termed as Load Balancing. There are two types of clustering architectures offered by Web Hosting services providers. It is a technique to group the users or web pages with similar characteristics. Any cluster should exhibit two main properties:

1. Low inter-class similarity
2. High intra-class similarity

Cluster analysis is used in a number of applications such as data analysis, image processing, market analysis etc. Clustering helps in gaining, overall distribution of patterns and correlation among data objects [1]. Clusters are created to provide greater scalability at lower price or better Availability (or both).

Different Clustering Algorithms for Web Opinions Development:

- Hierarchical clustering algorithm
- K-means clustering algorithm
- Density Based Clustering algorithm
- Partition clustering algorithm
- Spectral clustering algorithm
- Grid based clustering algorithm
- Incremental Clustering
- Scalable Distance Based Clustering

II. Architecture for Web Opinion Analysis

The special properties of Web opinions that do not exist in regular documents include the following:

- The content of messages is less focused.
- The messages are usually short in length ranging from a few words to a couple paragraphs.
- The language used in the messages is sparse because different users may use different terms to discuss the same topic.
- The messages contain many unknown terms or slang that do not exist in typical dictionary or ontology, e.g., iPhone and Xbox.
- There are many noises like unrelated text or typographical error so that many Web opinions do not fall into any categories.
- The volume of Web opinion messages is huge and ever expanding in an enormous Rate.
- The topics in these messages keep evolving. These different properties do not exist in typical documents.

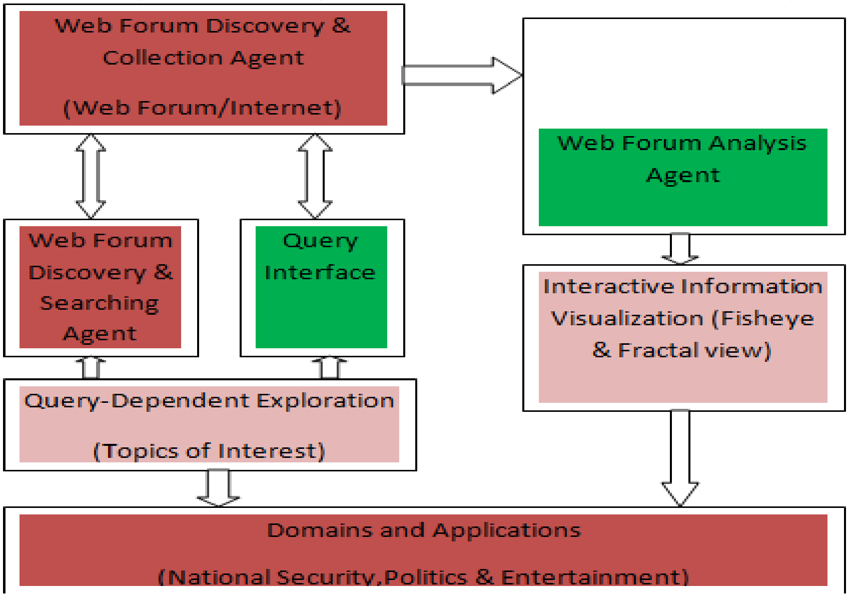

Web Forum Discovery & Collection Agent
(Web Forum/Internet)
Web Forum Analysis Agent
Web Forum Discovery & Searching Agent
Query Interface
Interactive Information Visualization (Fisheye & Fractal view)
Query-Dependent Exploration
(Topics of Interest)
Domains and Applications
(National Security,Politics & Entertainment)

Fig. 1 Architecture for Web opinion Analysis and Discovery

Here the Alert Agent plays a very important role for effective exploration of web opinions from the web. We present Web opinion clustering and information visualization techniques, which are components of a Web opinion analysis and understanding [5].In this Architecture has three major components. In the first component, i.e., Web forum discover and collection, a monitoring agent monitors a forum, and a web crawler fetches messages in the forum according to the hyperlink structure. The collected messages are analyzed with the importance on these three dimensions: member identity, timestamp of messages, and structure of threads. In the second component, i.e., Web forum content and link analysis, web utilizes machine learning and social network analysis techniques to extract useful knowledge. In the third component, i.e., user interface and interactive information visualization, we provide a user interface for users to submit their queries and present results through interactive visualization techniques for users to explore the forum social networks and content. Unlike Web or regular documents, Web opinions are usually less organized, short, and sparse text messages. Thus, traditional ways of clustering Web opinions become very challenging.

III. Related Work

A. Different Clustering Algorithms

1) K-Means and EM Algorithm

The K-Means is a prototype based clustering Algorithm. They select K initial centroids. For each point it finds closest centroids and assigns that point to the centroid. This forms a K cluster. The K- means is a partitional clustering algorithm. This is fast for low dimensional data and also it finds the exact sub clusters if large numbers of clusters are specified. We cannot use this in clustering web opinion because this cannot handle data of varying size and densities. More over the outliers are not identified and this is more cramp to the data which has a centroid. The web opinions keep on increasing and more over they don't rely on the centroid.

Donald Rubin proposed the expectation maximum algorithm. An expectation-maximization (EM) algorithm is for discovering the maximum likelihood or maximum a posterior (MAP) which estimates the parameters in statistical models. This model is for unobserved latent variables. EM is an iterative method. The EM algorithm finds the expectation of the log likelihood which is evaluated using the current estimate of parameters and a maximum step for finding the parameters maximized during the expectation of the log likelihood. This method requires the predefined set for clustering. But in web opinion this not quite possible since one cannot predict the number of clusters as they keep on growing.

2) SNN Algorithm

The shared nearest neighbor (SNN) is the enhanced method for density based clustering. The SSN and DBSCAN differ in the definition of the similarity between points in pairs. [5]SNN defines the similarity of the points in the pairs as the number of nearest neighbors the two points share. The density is measured by the sum of the similarities of the nearest neighbors of the point. Points that have high density are selected as the core points, and that with the low density are identified and removed as the noise. All the points similar to the core point are cluster together. They form cluster in elongated shapes so that they can overcome the different problems that occur in DBSCAN algorithm. It was also reported that SNN outperformed DBSCAN in a number of data set [7].SNN may achieve higher performance, it is not preferred in clustering Web opinions due to the fast growing number of Web opinions in social media.

3) Density-based spatial clustering of applications with noise

DBSCAN is a density-based cluster algorithm that can discover the clusters and filter the noise into a spatial database [6].The DBSCAN finds the number of clusters which starts from the predictable density distribution of parallel nodes. This consists of two parameters eps and minimum points required to form a cluster. This method starts with an arbitrary point that has never been visited. Once the neighborhood is retrieved and if it contains sufficient points, the clustering process is started. The density in the noisy region is lower than the density in the normal regions. [10] The boundary is known as the eps and the minimum eps is the neighborhood radius.

The process of determining clusters and finding density-reachable points repeats until all points in the data set are examined. Because the process of expanding a cluster or merging clusters in DBSCAN relies on the density reachability mechanism, some of the resulting clusters may be in non-convex or elongated shape. DBSCAN tends to merge clusters together and include more noisy threads in the clusters. The Limitation of DBSCAN is that it requires a predefined set of clusters which is not possible in web opinion. So the DBSCAN is not applicable for web opinion.

4) Scalable Distance Based Clustering (SDCA) Algorithm

The scalable density based clustering doesn't require any predefined set for clustering and this removes noise in a better way. This is illustrated using solid clusters in the initial step. When the size of the cluster keeps on increasing they are represented using the dotted lines. For every iteration a new circle in the dotted format grows. The points that are density reachable directly are not included since their size keeps on increasing. The points in the cluster are close to one another with a reasonable distance and this is not valid to the direct density reachable.

SDC Algorithm

```
1. Create instances of all points as unclassified points S={s1,s2……sn}
2. Repeat
3. Randomly select a point Pi in S
4. The number of points in eps- neighbourhood of Pi ≥ minpts
5. Create the initial cluster Cj by including the eps-neighbourhood points
6. S = S - Cj
```

```
7.  Else Pi is classified as X
8.  S = S – Cj
9.  Until S = ϕ
10. For each initial cluster Cj
11. Repeat
12. Find the centroid
13. eps = eps - ▲eps
14. Insert points from X in which the distance from the centroid of the cluster is larger than eps
15. Until no other points are found
16. The points that are found in X are considered as noise
```

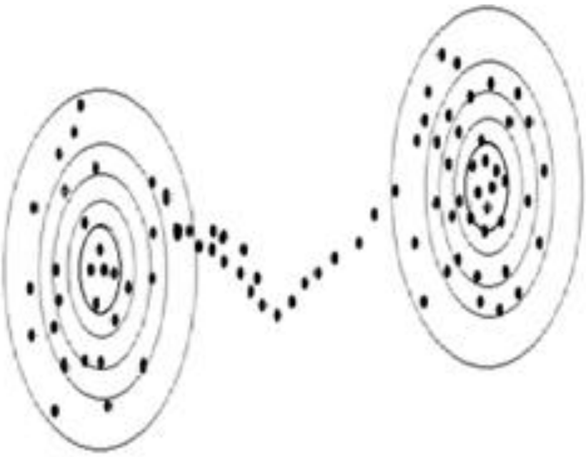


Fig. 2 Web Opinions Clustering by Scalable Distance based Clustering

The micro-accuracy and macro-accuracy are used as the metrics to measure the performance of SDC and benchmark with the performance of DBSCAN. Micro-accuracy measures the overall average clustering accuracy, whereas macro-accuracy measures the average of the clustering accuracy of all clusters.

$$microaccuracy = \frac{\sum_{i=1}^{|C|} |H_i|}{N} \quad (1)$$

$$macroaccuracy = \frac{\sum_{i=1}^{|C|} |H_i|/|C_i|}{|C|} \quad (2)$$

where |*C*| is the number of clusters created, |*Hi*| is the number of threads that is correctly classified in the cluster *Ci*, |*Ci*| is the number of threads in the cluster *Ci* and |*Ci*| is greater than one, and *N* is the total number of threads. DBSCAN and SDC do not require specifying the number of clusters to be formed. As a result, changing *eps* and *MinPts* will also affect the number of clusters being generated in addition to accuracy. The experiment shows that the total number of clusters decreases when *eps* increases. Similarly, the number of valid clusters decreases when *eps* increases. However, the difference between the total number of clusters and the number of valid clusters decreases when *eps* increases. That means that the number of invalid clusters decreases when *eps* increases. The total number of clusters and the number of valid clusters are about the same when *eps* is higher than 0.16.

SDC is also able to increase the size of a cluster gradually without including many noisy threads. It shows that the overall performance of SDC is substantially better than that of DBSCAN. A T-test shows that the difference between the performances of DBSCAN and SDC is significant at 0.05 level. Such finding ensures that a scalable distance-based approach is more suitable than a pure density-based approach for Web opinion clustering. Although SDC achieves good performance in clustering Web opinions, it has its own limitations.

- SDC does not require a predefined number of clusters, but it has two parameters *eps* and *MinPts* as inputs.
- The difference between micro-accuracy and macro-accuracy of SDC is not considerable.
- *eps* and *MinPts* are important in identifying the initial clusters.
- A systematic tuning of these two parameters is needed to achieve optimal performance.
- These parameters have impacts on micro- and macro-accuracy as well as the number of identified clusters.
- The tuning can be attuned to achieve the performance objectives.

B. Re-Ranking Schemes For Opinion Mining

The ranking is the core concept for clustering. There are several ranking schemes are available for Opinion Mining. After clustering each ranked list, we obtain a group of clusters each of which contains more or less relevant documents. By re-ranking, we expect to determine reliable clusters and adjust the relevance score of documents in each ranked list such that the relevance scores become more reasonable. To identify reliable clusters, we assign each cluster a reliability score. According to the Fusion Hypothesis, we use the overlap between clusters to compute the reliability of a cluster. There are several ranking methods for ranking an XML page. The most popular of these are the MaxEnt Ranking and p-norm Push ranking.

1) **Page Ranking:** Page rank algorithm use web page link structure to assign global importance to web pages. This uses the random web surfer method which refers to the process of starting at a random page and goes to through the link with uniform probability. The page rank has dynamic versions. [9] The versions are personalized web page rank and object rank methods.
2) **Object Ranking:** Perform keyword search in data base. This method uses the query term posting list as a set of random walk starting point and this continues the walk on the graph. The high recall search method is used. This requires the multiple iterations over all the nodes and this links the entire data base. The object rank has two operating modes such as the online mode and off line mode. [6] The online ranking mode performs the process once the query is given. This increases the retrieving process time for the given query. The offline ranking mode performs the pre computation for the top k results in the advance. This method is expensive and requires lots of storage space. This is not feasible for all the terms in the data dictionary.
3) **P-Norm Push Ranking Method:** The third algorithm we have tried is a general boosting-style supervised ranking algorithm called p-Norm Push Ranking. We describe this algorithm in more detail since it is quite new and we do not expect many readers to be familiar with it. The parameter "p" determines how much emphasis (or "push") is placed closer to the top of the ranked list, where p≥1. The p-Norm Push Ranking algorithm generalizes RankBoost (take p=1 for RankBoost). When p is set at a large value, the rankings at the top of the list are given higher priority (a large "push"), at the expense of possibly making misranks towards the bottom of the list.

Since for our application, we do not care about the rankings at the bottom of the list (i.e., we do not care about the exact rank ordering of the bad hypotheses), this algorithm is suitable for our problem. There is a tradeoff for the choice of p; larger p yields more accurate results at the very top of the list for the training data. If we want to consider more than simply the very top of the list, we may desire a smaller value of p. Note that larger values of p also require more training data in order to maintain generalization ability (as shown both by theoretical generalization bounds and experiments).

IV. Proposed Algorithm

The MaxEnt Ranking algorithm is a type of re-ranking algorithm. It prevents the occurrence of unrelated terms in the searches and it improves the efficiency of web opinions when compared to the p-norm push method. MaxEnt method is more effective in ranking and is more accurate.

A. Sampling and Pruning

Maximum Entropy models are useful for the task of ranking because they compute a reliable ranking probability for each hypothesis. We have tried two different sampling methods – single sampling and pair wise sampling. The first approach is to use each single hypothesis *hi* as a sample. Only the best hypothesis of each sentence is regarded as a positive sample; all the rest are regarded as negative samples. In general, absolute values of features are not good indicators of whether a hypothesis will be the best hypothesis for a sentence; in pair wise sampling we used each pair of hypotheses *(hi, hj)* as a sample. The value of a feature for a sample is the difference between its values for the two hypotheses. However, considering all pairs causes the number of samples to grow quadratically (O (N2)) with the number of hypotheses, compared to the linear growth with best/non-best sampling. To make the training and test procedures more efficient, we prune the data in several ways. We perform pruning by beam setting, removing candidate hypotheses that possess very low probabilities from the HMM, and during training we discard the hypotheses with very low F-measure scores. We also discard the pairs very close in performance or probability.

B. Decoding

If *f* is the ranking function, the MaxEnt model produces a probability for each un-pruned "crucial" pair: *prob (f (hi, hj) = 1)*, i.e., the probability that for the given sentence *hi* is a better hypothesis than *hj*. We need an additional decoding step to select the best hypothesis. Inspired by the caching idea and the multi-class solution proposed by (Platt et al. 2000), we use a dynamic decoding algorithm with complexity O (n) as follows. We scale the probability values into three types: Compare Result *(hi, hj)* = "better" if *prob (f (hi, hj) =1)* >δ1, "worse" if *prob(f(hi, hj) = 1)* <δ2, and "unsure" otherwise, where δ1≥δ2. 4

```
Prune
for i = 1 to n
Num = 0;
for j = 1 to n and j≠i
If CompareResult (hi, hj) = "worse"
Num++;
if Num>β then discard hi from H
Select
Initialize: i = 1, j = n
while (i<j)
if CompareResult(hi, hj) = "better"
discard hj from H;
j--;
else if CompareResult(hi, hj) = "worse"
discard hi from H;
i++;
else break;
```

V. Experiment Results

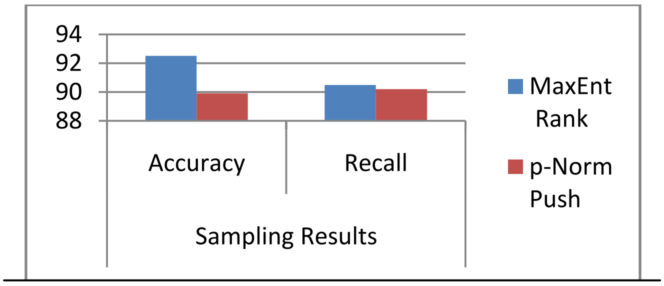
A. Effect of Pairwise Sampling

We have tried both single-hypothesis and pairwise sampling in MaxEnt-Rank and p-Norm Push Ranking. Table 5.1 shows that pairwise sampling helps both algorithms. MaxEnt-Rank benefited more from it, with precision and recall increased by 2.6% and 0.3% respectively.

TABLE I
Effect of Pairwise Sampling

Model	Sampling Results	
	Accuracy	Recall
MaxEnt Rank	92.5	90.5
p-Norm Push	89.9	90.2

Bar Chart 1. Comparative Analysis of MaxEnt and p-Norm Push Ranking Methods



Bar Chart displays the result of MaxEnt and p-Norm Push Ranking using Accuracy and Recall Parameters, in which the Accuracy and recall level are higher in MaxEnt Re-ranking method. To improve the Accuracy level of SDC Algorithm is combined with MaxEnt Ranking Methods.

VI. Conclusion

We concluded that the scalable distance based clustering method is the better method for web opinions Development. The advantage of SDC is that it groups the less relevant clustering into small groups when they are density- reachable. The limitations of SDC are only limited macro and micro accuracy is achieved. To improve the accuracy more, one can perform the involvement along with the clustering. Here, we associate MaxEnt Ranking Methods with SDC for improving the Accuracy level of web opinion Clustering .When compared with p-norm Push Ranking method, the MaxEnt Ranking Method ranks only the relevant terms and omits the unnecessary terms. Due to this, the accuracy is increased and is more effective.

References

[1] S. Banerjee, K. Ramanathan, and A. Gupta, “Clustering short texts using wikipedia,” in *Proc. ACM SIGIR*, Amsterdam, The Netherlands, 2007, pp. 787–788.

[2] B. Bicici and D. Yuret, “Locally scaled density based clustering,” in *Proc. ICANNGA*, 2007, pp.

[3] D. Bollegala, Y. Matsuo, and M. Ishizjka, “Measuring semantic similarity between words using Web search engines,” in *Proc. Int. WWW Conf.*, 2007, pp. 757–766.

[4] S. Das, A. Abraham, and A. Konar, “Automatic clustering using an improved differential evolution algorithm,” *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 1, pp. 218–237, Jan. 2008.

[5] Eugene Charniak and Mark Johnson. “Course-to- Fine N-Best Parsing and MaxEnt Discriminative Reranking”. *Proc. ACL2005*. pp. 173-180. Ann Arbor, USA

[6] David Chiang. ”A Hierarchical Phrase-Based Model for Statistical Machine Translation”., *Proc. ACL2005*. pp. Ann Arbor, USA

[7] L. Ertöz,M. Steinbach, and V. Kumar, “Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data,” in *Proc. 2nd SIAM Int. Conf. Data Mining*, San Francisco, CA, 2003.

[8] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, “Structure and evolution of blogspace,” *Commun. ACM*, vol. 47, no. 12, pp. 35–39, Dec. 2004.

[9] Q. Mei, C. Liu, H. Su, and C. Zhai, “A probabilistic approach to spatiotemporal theme pattern mining on weblogs,” in *Proc. Int. WWW Conf.*, Edinburgh, U.K., 2006, pp. 533–542.

[10] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz, “Why we blog,” *Commun. ACM*, vol. 47, no. 12, pp. 41–46, Dec. 2004.