ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPINION MINING AND SUMMARIZATION OF COMPANY REVIEWS

[1]KRITHIKA L.B , [2]ANAND MAHENDRAN
[1]School of Information Technology & Engineering, VIT University, Vellore, India
[2] School of Computer Science & Engineering, VIT University, Vellore, India
E-mail: [1]krithika.lb@vit.ac.in ,[2] manand@vit.ac.in

ABSTRACT

In the age of information any process begins with searching for information. Searching has become an unending syndrome. We search for information on vast topics and the frequent few topics are on product, opinion about company, education institution and news over the internet. Search results include indexed data from review forum, blog, tweet, emotional comment on websites and wiki. Data required for information lays all over flooded and piled with mix of unique, duplicate, custom generated, false portrait, broken sentence and multiple language. These data are unauthentic, meaningless and noisy. This paper work focus on extracting review comment data of company from review forums, reduce noise, group them and give meaning full summarized information and cumulative review rating of company's features to the user.

Keywords: *Natural Language Processing, Opinion Mining, Feature Based Summarization, User Review, Clustering*

1. INTRODUCTION

Data avalanche is an unmanageable attribute of the information age. We use search engine to search the data avalanche is a part of daily routine [1]. We get quick data form different mean which includes blogs, review, comments, tweets, wiki and information portals [2].These data available widely are generated from different source such as persons writing personal blog, review comment and tweets personally out of their own interest. There are big companies that get quick review because of large employee strength [3].Growing companies has made a practice of using Search Engine Optimization to add customized comment and review to add false propagation of their company on the internet [4].There are also automatic programs which spam review and content over the internet forums and review sites [5][6]. Extending further these reviews written contains information directly and indirectly about the company [7][8].Validation information out of these data cannot be achieved in isolation or small numbers, we require large data to be classified, grouped and noise reduced to extract the real information [9].

In our paperwork, we broadly categorize how data are extracted from company review portal, selection of top features and summarization of user review about company. Here our summarization is broadly based on top five features a company should posses. These features are obtained by ranking on the feature extracted [10].Feature selection in our work is different from method proposed for mining feature option by Hu, Minqing, and Bing Liu.2004 [11].Our work feature selection is based on summarization summation and ranking of feature extracted from review portal.

Feature: High compensation – Rating 4/5

Positive($ReviewRate_1$,…,$ReviewRate_n$) - Negative ($ReviewRate_1$,…, $ReviewRate_n$)

Feature: Flexible work time – Rating 3/5

Positive($ReviewRate_1$,…,$ReviewRate_n$) - Negative ($ReviewRate_1$,…, $ReviewRate_n$)

Summation: Infer Review Rate = Positive Review Rate – Negative Review Rate

Let us demonstrate review of IT Company to illustrate without review summarization and review summarization.

A. Without summarization and Feature grouping

User can have review of only one feature reviewed at a given time and summarization is not possible. One user can review (R_1) of Feature (F_1)

at Instance I ,which can be either a positive or a negative review.

$$User\ Review\ =(F_1\ *\ R_1\ )\ *I \qquad (1)$$

Assume it takes t_1 time for I_1 ,then

$$T(I_1) = t_1 \qquad (2)$$

where,

T is the time taken for Instance

I_1 is one Instance

and t_1 is the time take for one review of one feature .

B. With summarization and Feature grouping

Using without summarization and feature grouping, N review of N features of a company will require t_n time for review and another t_n to summarize the reviews.

Time taken for N review without our method would result in time which is given by,

$$T\ (I_n\ )=t_{\ n*n} \qquad (3)$$

With our implementation (i.e) with summarization and feature grouping in the same time $T(I_1)$ our method get abstract review of N feature of N commentator.

$$N\ Feature\ UserReview = N*F_{(n)}*R_{(n)}*I \qquad (4)$$

where,

I is the time required to read $R_{(n)}$review of $F_{(n)}$ features about a company by N commentator.

Our paper summarizes the review and gives the following,

Feature: High compensation – Rating 4/5

Positive(ReviewRate$_1$,…,ReviewRate$_n$) - Negative (ReviewRate$_1$,…, ReviewRate $_n$)

Feature: Flexible work time – Rating 3/5

Positive(ReviewRate$_1$,…,ReviewRate$_n$) - Negative (ReviewRate$_1$,…, ReviewRate $_n$)

With our work User can review N feature of N review summarized at I instance

$$\Sigma S = \{F_{1\,R}\ , F_{2\,R,}\ F_{3\ R}, F_{4\,R}, F_{5\ R}\ \} \qquad (5)$$

where,

S is the summarized review

F_n are features

R is the review

The strength of the opinion on a review given by the user is emphasized based on the scale value added to the comment/review. The scales largely vary from verbal scale as *good, bad, very good, very bad, and normal*. Numerical scales like whole numbers and percentage. Character scales like ☺ , ☹and jargon words like *ok, oops*. Process the scales for flood of data is quite difficult, where we require transformation and loading into single format to match data different sources. This paper work considers only ranking based on numerical whole numbers in the experiment conducted.

2. RELATED WORK

K- means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean[12] .The term "k-means" was first used by James MacQueen in 1967[13]. Though the idea goes back to Hugo Steinhaus in 1957[14].In 1965, A more efficient version was proposed and published in Fortran by Hartigan and Wong in 1979[15].

Forgy method: The Forgy method randomly chooses k observations from the data set and uses these as the initial means[16]. The random partition method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial mean to be the centroid of the cluster's randomly assigned points. The Forgy method tends to spread the initial means out, while random partition places all of them close to the center of the data set. According to Hamerly et al.,[17] the random partition method is generally preferable for algorithms such as the k-harmonic means and fuzzy k-means. For expectation maximization and standard k-means algorithms, the Forgy method of initialization is preferable.

3. PROPOSED TECHNIQUE

Our approach is to first find top five feature of company which the user uses to describe and rate the company. For a large set of data taken from company review portal using k-mean cluster and Forgy method we come out with top five features of the companies that are emphasized by user in the review portals. These five features is a subset of many features that were discussed. Then we segregate the review to fall into this subset. We capture the min, max and mean of summation of all the rating about the feature of a company [18].

A. *Data extraction*

This process is where data required for process are extracted from glassdoor.com using custom build java bot. The extracted information from glass box portal are purely used for experimental and educational purpose only. No data or result would be commercialized [19].

Our work require an automatic program to extract data, below is the pseudo code:

```
locate page
begin:parse page
locate rating(get value for xpath)
if(word = = pros || con || suggestion)
{match found
extract start to end of line
}while != eop
end:store data
```

B. *Data cleansing*

In data cleansing, we clean the data by removing the maximum and minimum threshold of keywords form the raw data. Less frequent used word labels and words without label are also truncated [20].

Sample record extracted by our program

Delivery Head— Rating 4.0 out of 5
Pros - Great Employer and high value system
Cons - Few people of company have made it very operational oriented

C. *Data transformation*

When data from multiple sources converge to form the raw data source then we employed data transformation where we segregate all the data to one common format. Segregation we remove unwanted field, or convert an attribute to a required format to match the requirement of the input data to our summarization engine [21].

Table:1 : Sample of extracted data

Extracted data	Rating	Average
management none add	4	0.089
2013 pros very good	4	0.119
place to start your	4	0.119
Good consultant	4	0.06
industry standards	4	0.06
this helpful — wed	4	0.119
jun 2013 pros very	4	0.119
higher management	4	0.06
23 project manager award	4	0.089
jul 2013 pros the	4	0.119
management focus	4	0.06
market standards	4	0.06
delivery manager	4	0.06
salary structure	4	0.06

D. *Data grouping*

Generally where large scattered data inferences are required, grouping make the large data into sets of data falling under specifies target groups. We employed forgy method to group the data into groups. Features are grouped based on top 10 features identified from extraction. Top feature identification is based on ranking among the selected 10 features [16].

E. *Data training*

Any machine learning experiment require, training set to train the machine to perform unsupervised actions. We employee labeled data for training [22].

F. *Data loading*

Large data after extraction, transformation, cleansing and grouping is now ready for upload to the machine for prediction.

4. EXPERIMENT & RESULT

Procedure:
Step 1 - Observation of review extracted are partitioned into k sets. Where, k set is lesser that or equal to n observation.
Step 2 - Cleaning & transformation.
We use an SQL procedure to truncate maximum and minimum label occurrence threshold.
Step 3 - Grouping & partition the observation made from the review and review rate. Partition is done using Forgy and Random Partition.
Step 4 - We adopt semi supervised feature classification by labeling the observation.
Step 5 - Map the data to corresponding set.
Step 6 - Summarization of review and chart horizontal box plot.
Step 7 - For every run the observation keeps increasing
Step 1 – 6 repeated, gives refined result every run.

A. *Extraction*

Running the java program written for our experiment can extract 5000 reviews for glassdoor.com in less than a minute. Figure-1 in appendix shows the work of review extraction,

B. *Cleaning & transformation*

This program is custom written for our experiment which follows the below mentioned logic. This code snippet is run for every cycle new data is update into our data store.

```
start:
parse data
match:rank || rank column
match:comment || comment column
truncate rank value “>5 , <0”
                || update rank column
truncate repeated phrase count threshold >3
                || update comment column
end:
```

C. *Ranking of features from the extracted data:*

Data grouping, from the extracted data and cleansed data we take the top 12 feature that most users describe about the company the write review. We have reviewed by grouping and choosing the top 5 features.In the table II, the features listed are used for grouping and feature rating. Figure 2 shows the features ranking based on occurrences in user review, data used in this graph is extracted and cleansed data,

Table 2: Features Listed In The Table Are Used For Grouping And Feature Rating

Feature Rating	Rank
Culture & Values	1
Management	2
Career Opportunities	3
Work/Life Balance	4
Comp & Benefits	5

D. *Synonyms and common interchangeable word tag for label*

Label is defined as a set of common interchangeable word tag .These word tag combination is used to categorize review under the label.

L{word tag 1 … word tag n }
Here ‘L’ represents the label.

Feature labels used in our paperwork are Culture, Value, Management, Career,Opportunities, Work/ Life balance, Compensation, and Benefits.

Culture { work culture , corporate culture , weekend bash , birthday party , team outing }

Value { worth , significance , respect }

Management { PM , boss , TL , manager , Management }

Career { Career , job , profession , opening , vacancy }

Opportunities { onsite , growth , promotion , learning }
Work / Life balance{maternity , parenthood , motherhood , leaves }

Compensation {salary , hike , remuneration , income }
Benefits { cab , dinner , remuneration , med claim , insurance , LIC , reclaims }

E. *Summarization and plotting*

We have the Observation Set (x_1, .., x_n).We then partition ‘n’ observation into k sets ;where, k < = n and k represents the features of the company

where,
Observation in our experiment n = 5000(review from web portals)

Partition, $K_5 = F_1$, F_2 ,F_3,F_4 and F_5

where,
F are the five features used in our paper using ...(5) we get below outcome.

Table 3 - Outcome of the paperwork.

Feature Rating	Average	0	1	2	3	4	5
Culture & Values	2		🟩	🟩	🟩		
Management	3			🟩	🟩	🟩	
Career Opportunities	2.5	🟩	🟩	🟩	🟩	🟩	🟩
Work/Life Balance	2.5		🟩	🟩	🟩	🟩	
Comp & Benefits	4.6				🟩	🟩	🟩

The Features in the table were top five features identified based on our work after extracting and ranking 5000 reviews from glassdoor.com. Review comments where again grouped under one of the five features and mean average is listed under average column. The highlighted bar represents box plot of each feature horizontally to depict the high, low and mean value of a feature [23].

5. CONCLUSION

Our work was collection of review about company, reduce the noise and summarize review based on top features. The work consolidates the cumulative review and summarizes it to the user so that user gets to read cumulative mean of review. Future work would include add more interchangeable word tag, extracting of data from all type of review rating as of now we have focused only on numerical whole number rating system.

REFRENCES:

[1] Lee Rainie,”Big jump in search engine use” available online: http://www.pewinternet.org/Reports/2005/Big-jump-in-search-engine-use/DataMemo.aspx

[2] Khare, Shashi Kant, Neelam Thapa, and K. C. Sahoo. "Internet as a source of information: A survey of Ph. D Scholars." Annals of Library and Information Studies 54.4 (2007): 201.

[3] Dias, Cláudia. "Corporate portals: a literature review of a new concept in Information Management." International Journal of Information Management 21.4 (2001): 269-287.

[4] Chen, Chen-Yuan, et al. "The exploration of internet marketing strategy by search engine optimization: a critical review and comparison." African Journal of Business Management 5.12 (2011): 4644-4649.

[5] Kreibich, Christian, et al. "On the Spam Campaign Trail." LEET 8 (2008): 1-9.

[6] Chu, Zi, et al. "Who is tweeting on Twitter: human, bot, or cyborg?." Proceedings of the 26th annual computer security applications conference. ACM, 2010.

[7] Kinnon, Rebecca. "China’s Censorship 2.0: How companies censor bloggers." First Monday 14.2 (2009).

[8] Kittur, Aniket, et al. "He says, she says: conflict and coordination in Wikipedia." Proceedings of SIGCHI conference on Human factors in computing systems. ACM, 2007.

[9] Kittur, Aniket, et al. "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie." World Wide Web 1.2 (2007): 19

[10] Feature Selection for Knowledge Discovery and Data Mining By Huan Liu, Hiroshi

[11] Hu, Minqing, and Bing Liu. "Mining opinion features in customer reviews." AAAI. . 2004

[12] Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". Journal of the Royal Statistical Society, Series C 28 (1): 100–108.

[13] MacQueen, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201.

[14] Steinhaus, H. (1957). "Sur la division des corps matériels en parties". Bull. Acad. Polon. Sci. (in French) 4 (12): 801–804. MR 0090073. Zbl 0079.16403.

[15] Lloyd, S. P. (1957). "Least square quantization in PCM". Bell Telephone Laboratories Paper. Published in journal much later: Lloyd., S. P. (1982). "Least squares quantization in PCM". IEEE Transactions on Information Theory 28 (2): 129–137. doi:10.1109/TIT.1982.

[16] E.W. Forgy (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics 21: 768–769.

[17] Hamerly, G. and Elkan, C. (2002). "Alternatives to the k-means algorithm that find better clusterings". Proceedings of the eleventh international conference on Information and knowledge management (CIKM).

[18] Lewis, David D. "Feature selection and feature extraction for text categorization." Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992.

[19] Laender, Alberto HF, et al. "A brief survey of web data extraction tools." ACM Sigmod Record 31.2 (2002): 84-93.

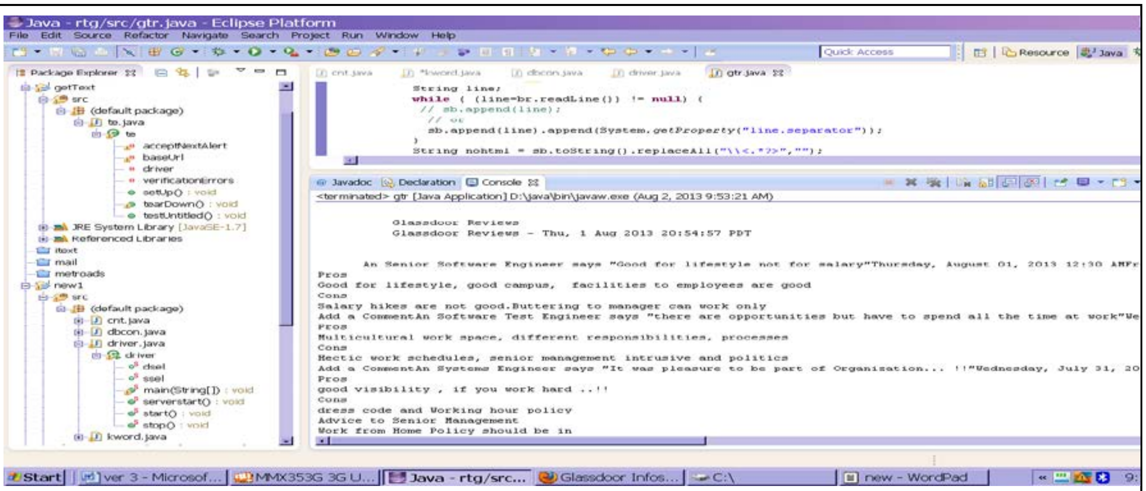
[20] Rahm, Erhard, and Hong Hai Do. "Data cleaning: Problems and current approaches." IEEE Data Eng. Bull. 23.4 (2000): 3-13.

[21] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.

[22] Lewis, David D. "Feature selection and feature extraction for text categorization." Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992.

[23] Benjamini, Y. (1988). "Opening the Box of a Boxplot". The American Statistician 42 (4): 257–262. doi:10.2307/2685133. JSTOR 2685133.

APPENDIX:



```
            String line;
            while ( (line=br.readLine()) != null) {
             // sb.append(line);
              // or
              sb.append(line).append(System.getProperty("line.separator"));
            }
            String nohtml = sb.toString().replaceAll("\\<.*?>","");
```

```
<terminated> gtr [Java Application] D:\java\bin\javaw.exe (Aug 2, 2013 9:53:21 AM)

          Glassdoor Reviews
          Glassdoor Reviews - Thu, 1 Aug 2013 20:54:57 PDT


      An Senior Software Engineer says "Good for lifestyle not for salary"Thursday, August 01, 2013 12:30 AMFr
Pros
Good for lifestyle, good campus,  facilities to employees are good
Cons
Salary hikes are not good.Buttering to manager can work only
Add a CommentAn Software Test Engineer says "there are opportunities but have to spend all the time at work"We
Pros
Multicultural work space, different responsibilities, processes
Cons
Hectic work schedules, senior management intrusive and politics
Add a CommentAn Systems Engineer says "It was pleasure to be part of Organization... !!"Wednesday, July 31, 20
Pros
good visibility , if you work hard ..!!
Cons
dress code and Working hour policy
Advice to Senior Management
Work from Home Policy should be in
```

Figure 1: Review extraction

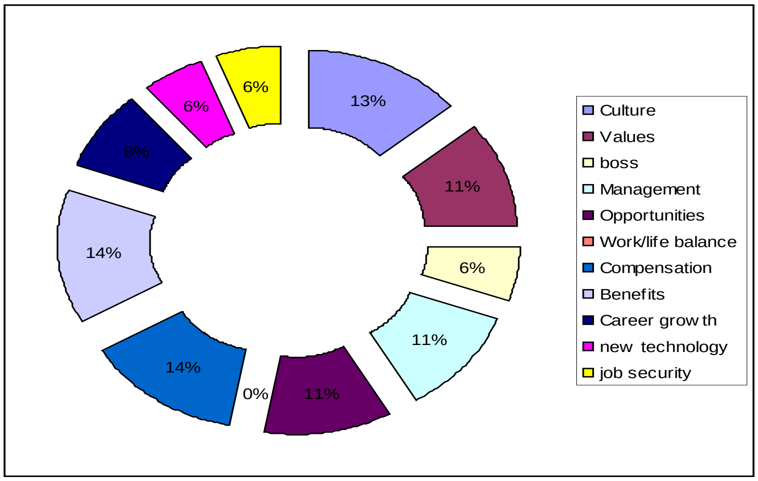


Figure 2: Ranking based on occurrences in user review, Data used in this graph is extracted and cleansed data