Book Recommendation System Using Opinion Mining Technique

Shahab Saquib Sohail
Dept. of Computer Science
Aligarh Muslim University, Aligarh,
U.P.
shahabsaquibsohail@gmail.com

Jamshed Siddiqui
Dept. of Computer Science
Aligarh Muslim University, Aligarh,
U.P.
Jamshed_faiza@rediffmail.com

Rashid Ali
Dept. of Computer Engineering
Aligarh Muslim University, Aligarh,
U.P.
rashidaliamu@rediffmail.com

Abstract—With the changing trends in technologies, and brisk growth in Internet, daily life of an individual has also changed at a very fast pace. The impacts of these technologies are so diverse that it has affected almost every sphere of the life. People start using applications of the Internet in their daily life; they prefer online shopping for their needs more and more. For academician, researchers and students, purchasing the desired book from huge collections of books on the Internet is very tedious work. In this paper, we presented a recommendation technique based on opinion mining to propose top ranked books on different discipline of the computer science. Based on the need of the customers and the reviews collected from them, we have categorized features for the books. We analyze the features on the basis of several characteristics that we have categorized and reviews of the users. Weights are assigned to categorized features according to their importance and usage, and accordingly the ranks are given. Finally, top ten ranked books are listed. This method is expected to be helpful for millions of the users who seek for desired books.

Keywords—opinion mining; recommendation technique; user review; feature analysis

I. INTRODUCTION

"Necessity is the mother of invention," the famous proverb is practically experienced in our daily life as we are growing and moving towards the advancement in technologies. These technologies are giving birth to the modern tools and techniques to fulfill our daily needs. Today a huge number of users are using the Internet. The developed countries like Germany and U.K have approximately 83% Internet user of their population, whereas China leads the overall contribution to the Internet user in the world, which counts to 22.4%. USA has 78.1% Internet users of their population, a contribution of 10.2% of overall users in the world [1]. This acceleration in the users of the Internet in recent days has changed the style people live, they think and they work.

With the changing trends in technologies, daily life of an individual has also changed at a very fast pace. People prefer online shopping for their needs more and more. To make online shopping easy and reliable a good number of product recommendation techniques are proposed by many researchers in last few years [2,3,4].

There are several known and frequent used techniques to recommend products, amongst which opinion mining is one of the emerging and efficient method being used worldwide.

Customers' reviews are the basis for opinion mining technique. Finding and summarizing the opinion from huge amount of reviews from the customers, is also very tedious for business.

For researchers, Opinion mining is a very hot topic in the field of data mining. The main issue to consider is to find (a) product feature and (b) analysis comments, whether positive or negative, as described in [5, 6, 7]. People generally use some pre-determined terms to interpret it as a positive or negative comments, like, better, good, nice, well written, highly recommend etc. are treated as positive terms and worst, time consuming, bad, not recommended, etc. are termed as negative comments. In [12], opinion retrieval is perceived as a two-step task, finding relevant documents and re-ranking these documents by opinion scores. The reviews are given by the human and it is very evident that to understand the review one should perceive it as human being. Finding the comments only is not sufficient. Sometimes things are different then what they seem to appear. Let us consider the following example:

"*I highly recommend this book for those who want to waste their time and money. If you are really sincere to get some knowledge into your bucket, another one is the better option*"

Though the sentence above has terms like highly recommend and better but both terms are being used in a negative sense for some specific book, keeping only the positive and negative aspects of the terms and processing on these basis is not sufficient alone to extract opinion for a better conclusion.

In this paper, we have recommended top ten (10) ranked books on different disciplines of computer science. The opinion mining techniques are applied on the customer reviews. The customer's reviews are obtained simply by common query terms to search engines for different books of various disciplines of computer science. Querying the books generates various books detail and opinion of these books can easily be known using query for reviews. These reviews are analyzed and checked with those features of books, which a user usually love to have. After using a ranking system, described in section 3, books are ranked in descending order of their ranking points.

Rest of the paper is organized as follows. In section 2 we gave a brief overview of the related work in the area of opinion mining, section 3 contains the working of the recommendation system, in section 4 experimental results are discussed, finally we summarize the conclusion in section 5.

Opinion Mining is one of the esteem branch of data mining which has been approached recently. We can find a detail survey in [8]. Opinions can be thought as a combination of words, sentiments or documents, an extensive work on these factors can be found in [9, 10, 11] respectively. However, some people believe that opinion is topic dependant and the above methods do not have consideration for this [12].

The TREC (Text REtrieval Conference) came out with a Blog Track with a major task of information retrieval [13]. In the similar way, NTCIR set up an Opinion Analysis task with multi lingual test beds in Chinese, Japanese and English [14].

Mining product review is one of the major focus for research in recent days [15]. There is a canvass on comparison for opinions [16]. In this work, we did not just compare the opinion but interpreted those opinions to extract features and rank the product based on the opinions. There are also good works discussed on document level classification in [17, 18, 19]. In some way, we have a similar approach to rank the books in this paper.

The study on sentence level classification is done in [4, 6, 21], our work differs from these analyses as we are working on related feature of the specified topic but a review sentence may be of multiple features. Even the orientation on these features may differ significantly. e.g., "the picture quality of the camera is very good but strength of its body is very weak", here "picture quality" and "strength of body" are features. Both the features have different types of opinions, one is having positive and other is negative. Related works on these topics include [20, 21, 22].

The reference [4] explores the clues for opinion mining and summarization. It makes use of lexicon-based method to find that expressed opinion for some specific product is positive or negative.

However, we have not automated our system and to best of our knowledge, we did not find any organized research work to recommend books on any specific field. This work is a specialized job combining various techniques to recommend the top books.

III.   PRODUCT RECOMMENDATION TECHNIQUE

In this section, first, we give architecture for our recommender technique followed by detailed explanation then the procedure and techniques are discussed.

A. Architecture of the Recommendation Technique

A block diagram is shown in the figure 1. As discussed in the previous sections, in this paper we intend to sort out the top computer books of various discipline of the computer science, therefore a ranked list of top 10 books is prepared. In this process, first we pass queries to search engines to obtain the names of the books of the discipline concerned. We selected more than 10 disciplines of the computer science, for every discipline we pass query like, "books on the 'discipline concerned' ", e.g., for books on cloud computing the query is, "books on cloud computing". The names of the books that appear in top 100 links, with other details like, Title, Author, publisher, edition, price and number of pages are stored in a separate file, for this, Search Engine Optimization (SEO) tools are used.

Again, we pass different queries for the books stored to get the user review. With the help of this, we perform task to find opinion. When the opinion extraction is complete, then, we start ranking the items (books in this case) on scores. On a specific book, we may have several types of comments; these comments are positive or negative. Comments may be neutral, that contains neither negative nor positive terms in reviews.

B. Review terms evaluation

An example for the different types of comments that we encounter while processing the opinion on books, is illustrated in table.1.

The review comments of the respective books are stored. Based on these reviews, we give weights to different common features, categorized based on general book features and opinion of the user to different books. The basis for the point scale is the number of opinions available and corresponding reviews from the users. The weights are assigned in multiple of 0.5 according to the opinions obtained and processed as described in table 1 below. The users may have reviews positive and negative both. The positive terms give more value to features and negative terms may lower the weights.

TABLE 1. POINT SCALE FOR REVIEW COMMENTS

Review/ Opinion terms	Sense	Points scale
Very useful, basic and very enthusiastic, great addition, invaluable as a reference guide, written by an experienced team, exactly what I needed, good job, covering everything you need to be aware of, especially appreciated, awesome book	Positive (+)	0.5 to 5
Worthwhile at perhaps half the price, real disappointment, Not a general discussion, fluffy, worst business book purchased, copied, biggest failing	Negative (-)	-0.5 to -5
Not bad, No review	Neutral	0

C. Product Feature Analysis

We categorize seven features to rank the books. The feature categorized and the weights given to respective features are tabulated below in Table 2. We consider W_{OC}, W_H, W_M, W_A, W_I, W_{OT}, and W_P as weights assigned to the features Occurrence, Helpfulness, Material, Availability, Irrelevancy, and Price respectively. The ranges of the weights are also tabulated in table 2.

Occurrence means the multiple appearances of the same books on single query. 'Helpfulness' and 'material' are concerned with the books review, If people find the book helpful and material inside the book relevant, the scores can be given up to 5 according to the review obtained. The scores are assigned in multiple of 0.5 according to the opinion obtained and the terms used in the review.
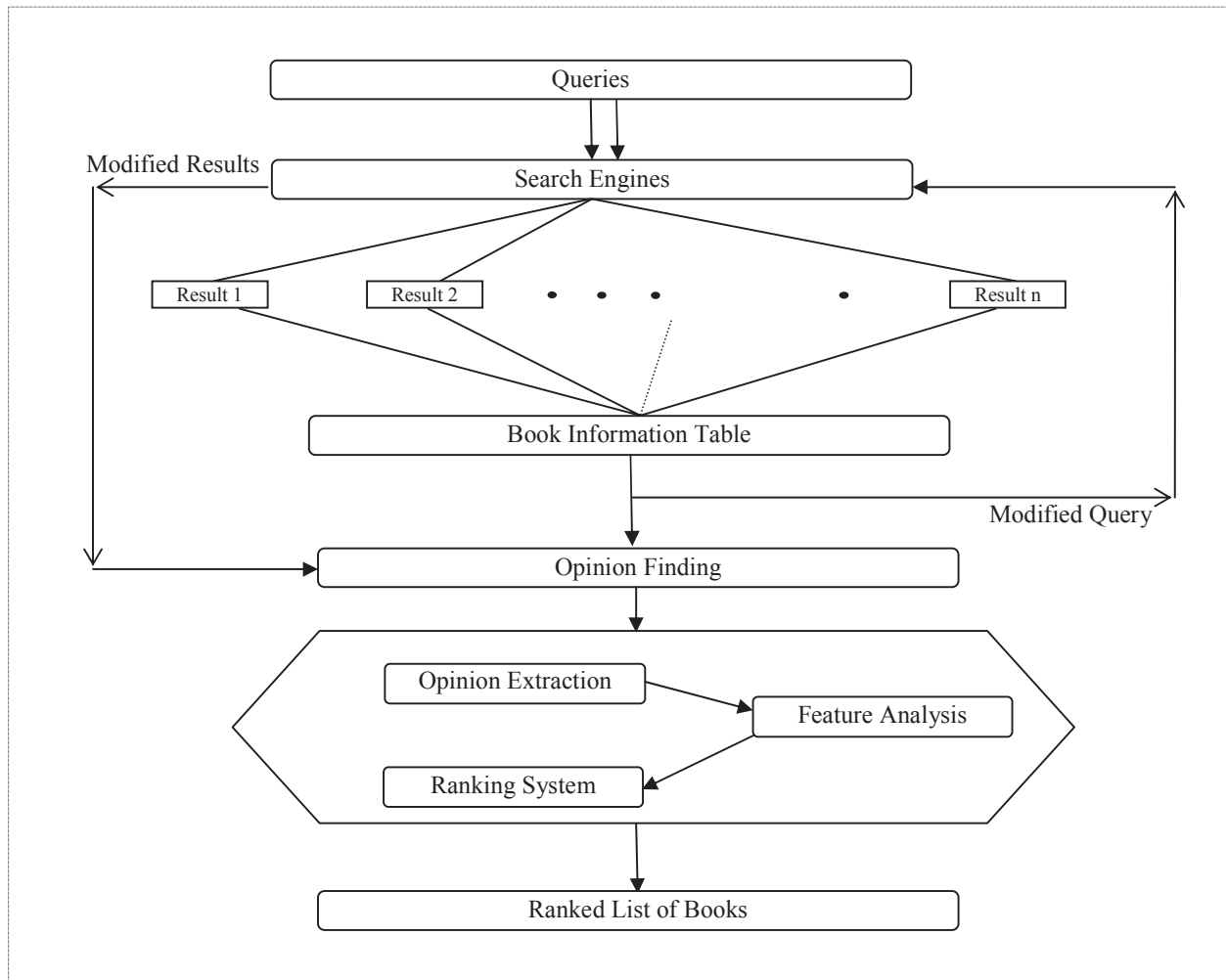
Fig 1. Architecture of the Recommender Technique

TABLE 2. FEATURES AND WEIGHTS

Features	Weights Assigned	Weight Range
Occurrence	W_{OC}	$0 \le W_{OC} \le 2$
Helpfulness	W_H	$0 \le W_H \le 5$
Material	W_M	$0 \le W_M \le 5$
Availability	W_A	$0 \le W_A \le 2$
Irrelevancy	W_I	$-5 \le W_I \le 0$
Price	W_P	$0 \le W_P \le 4$
Others	W_{OT}	$-2 \le W_{OT} \le 2$

Sometimes a book is well written but it is meant for some specific group of people and it is not going to be helpful for masses.

However, the material it contains is valuable, e.g., a book written for Master degree engineering students can never be helpful for first year Bachelor students of engineering, but the material inside the books may be valuable and well organized.

Availability is a non-technical feature, which tells the market value, edition of the book and its publishing services etc.

In most of the opinions, readers use to say about some book that content of the book is irrelevant or book seems to club everything together, etc. these types of Irrelevancy is categorized as a different feature and it is given weight in negative value to predict the rank in more precise and accurate measures. It is the case with another feature 'price', if the cost of the book is too much and customers had a complaint against it, the rank of the book will be affected.

Table 3. Scores based on weight distributed to products features

Products	W_{OC}	W_H	W_M	W_A	W_{OT}	W_I	W_P	T	NS
P1	0.5	1	0.5	0	0	0	1	3	0.15
P2	**1**	**3.5**	**3.5**	**2**	**2.5**	**-1.5**	**-0.5**	**10.5**	**0.525**
P3	0	2	2.5	1.5	2	-1	0	7	0.35
P4	1	0.5	0.5	0.5	1	-1.5	-1	1	0.05
P5	0	2.5	2.5	0.5	1.5	0	0	7	0.35
P6	1	1.5	1.5	0.5	1	-0.5	0	5	0.25
P7	**1**	**3**	**3.5**	**1**	**2**	**-1.5**	**-0.5**	**8.5**	**0.425**
P8	0.5	2	1	0.5	1.5	0	0	5.5	0.275
P9	**1**	**3**	**3**	**1**	**1.5**	**-1**	**0**	**8.5**	**0.425**
P10	0.5	3	2	0.5	1	-2.5	0	4.5	0.225
P11	0	2	1	0.5	0.5	0	0	4	0.2
P12	0	3	2	1	1	-0.5	0	6.5	0.325
P13	**0**	**4**	**3**	**2**	**2**	**-0.5**	**0**	**10.5**	**0.525**
P14	0	1	0.5	1	1	-1	0	2.5	0.125
P15	**1**	**3**	**3**	**1**	**2**	**-0.5**	**0**	**9.5**	**0.475**
P16	0.5	3	2	1	1	0	0	7.5	0.375
P17	0.5	2	2	1	1	-1	0	5.5	0.275
P18	**1.5**	**3**	**3**	**1**	**1**	**0**	**0**	**9.5**	**0.475**
P19	0	0	0	1	0	0	0	1	0.05
P20	1.5	2	2	1	1	-1	0	6.5	0.325
P21	1	4	3	1	2	-1.5	-1.5	8	0.4
P22	**1.5**	**3**	**4**	**1**	**2**	**0**	**0**	**11.5**	**0.575**
P23	1	2	3	1	2	-1	0	8	0.4
P24	0.5	0	0	1	0	0	0	1.5	0.075
P25	1	1.5	2.5	1.5	1.5	0	0	8	0.4
P26	0.5	1	0	1	2	-1	-1	2.5	0.125
P27	0	0	0	1	0.5	0	0	1.5	0.075
P28	**0.5**	**2.5**	**3.5**	**1**	**2**	**0**	**0**	**9.5**	**0.475**
P29	**0.5**	**2.5**	**3**	**1**	**1.5**	**0**	**0**	**8.5**	**0.425**
P30	0.5	0	0	1	1	0	0	2.5	0.125
P31	0	2	3.5	1	1.5	0	0	8	0.4
P32	0	0	0	1	1	0	0	2	0.1
P33	0	0	0	1	1	0	0	2	0.1
P34	0	1.5	2.5	1	1	0	0	6	0.3
P35	**1**	**3.5**	**3.5**	**1**	**2.5**	**0**	**0**	**11.5**	**0.575**
P36	**0.5**	**3**	**3.5**	**1**	**2**	**-1**	**0**	**9**	**0.450**
P37	1	0	0	1	1.5	0	0	3.5	0.175
P38	1	0	0	1	1	0	0	3	0.15
P39	1	0	0	1	1.5	0	0	3.5	0.175
P40	0	2	3	1	2	0	0	8	0.4
P41	0.5	2.5	3	1	1.5	-0.5	0	8	0.4
P42	1	0	0	1	1	0	0	3	0.15
P43	1	0	0	1	1	0	0	3	0.15

Table 4. Ranked list of products with respective scores

Products	Occurrence	helpfulness	material	availability	others	irrelevancy	price	total	Normalized
P22	1.5	3	4	1	2	0	0	11.5	0.575
P35	1	3.5	3.5	1	2.5	0	0	11.5	0.575
P2	1	3.5	3.5	2	2.5	-1.5	-0.5	10.5	0.525
P13	0	4	3	2	2	-0.5	0	10.5	0.525
P15	1	3	3	1	2	-0.5	0	9.5	0.475
P18	1.5	3	3	1	1	0	0	9.5	0.475
P28	0.5	2.5	3.5	1	2	0	0	9.5	0.475
P36	0.5	3	3.5	1	2	-1	0	9	0.450
P7	1	3	3.5	1	2	-1.5	-0.5	8.5	0.425
P9	1	3	3	1	1.5	-1	0	8.5	0.425
P29	0.5	2.5	3	1	1.5	0	0	8.5	0.425

Table 5. Final Ranked List of Books

Title	Author	Publisher	Year	Edition	Price	Page
Cloud Computing Explained: Implementation Handbook for Enterprises	John Rhoton	Recursive Press	2010	2	1,644	472
Microsoft Private Cloud Computing	Aidan Finn, Hans Vredevoort, Patrick Lownds, Damian Flynn	Sybex	2012	1	1,733	408
Cloud Computing: Web-Based Applications That Change the Way You Work and Collaborate Online	Michael Miller	Que Publishing	2008	1	990	312
THE CLOUD AT YOUR SERVICE	Jothy Rosenberg, Arthur Mateos	Manning Publications	2010	Pap/Psc	1,101	200
Enterprise Cloud Computing: A Strategy Guide for Business and Technology Leaders	Andy Mulholland , Jon Pyke , Peter Fingar	Meghan-Kiffer Press	2010		1,664	260
CLOUD SECURITY AND PRIVACY: AN ENTERPRISE PERSPECTIVE ON RISKS AND COMPLIANCE	Tim Mather, Subra Kumaraswamy, Shahed Latif	O'Reilly Media	2009	1	1,562	338
Cloud Computing for Libraries (THE TECH SET® #11)	Marshall Breeding	ALA	2012		3,225	121
Executive's Guide to Cloud Computing	Eric A. Marks, Bob Lozano	Wiley	2010	1	1,904	304
Cloud Application Architectures: Building Applications and Infrastructure in the Cloud	George Reese	O'Reilly	2009	1	1,009	208
Cloud Computing Bible	Barrie Sosinsky	Wiley	2011	1	1,496	532
Cloud Computing: Automating the Virtualized Data Center	Venkata Josyula; Malcolm Orr; Greg Page	Cisco Press	2011	1	1,593	336

W_{OT} refers to the features other than the above six described category, that deals with issues like how many users have rated the product? And how many stars have been rated?, how is the book cover and prints?, whether this particular book is referred by the customers of the other book as well?, checking the biasness in opinion, and referring other book as best while comparing, etc.

D. Ranking Procedure

We find out the total Opinion score for each item with the help of the details provided in table 2.

Let 'T' is the Total Opinion score for each item,

$$T = W_{OC} + W_H + W_M + W_A + W_I + W_{OT} + W_P \text{-------------- (1)}$$

If, $M = W_{OCMAX} + W_{HMAX} + W_{MMAX} + W_{AMAX} + W_{IMAX} + W_{OTMAX} + W_{PMAX}$ ------------------------------- (2)

Where W_{OCMAX}, W_{HMAX}, W_{MMAX}, W_{AMAX}, W_{IMAX}, W_{OTMAX}, and W_{PMAX} are the maximum value assigned to the weights of respective features.

We obtain Normalized Opinion score 'NS' = T/M ------ (3)

Finally we sort top books having maximum normalized scores in descending order of scores, thus we obtain top ten ranked books of a specific topic.

IV. EXPERIMENTAL RESULTS

Since we are recommending books, we did not take books name from any bookstall instead we searched the books for different disciplines of the computer science in search engines like Google. These books are then again searched with a modified query to get the review of the books and to obtain opinion of the customers on these books. We perform feature analysis based on the scores and weights given to various features categorized in the table 1 and table 2.

A. Scoring Table

We performed our task for more than 10 different disciplines of the computer science, here in this paper we intend to show the result for one book. We have chosen cloud computing for the purpose. For first 100 results on Search Engine Result Page (SERP) for the query, "*books on cloud computing*" we get 43 different books name. In addition, we repeat the process for all these 43 different books. The scores for all the books after processing their features and assigning weights to them, as described in section 2, is shown in the table 3.

B. Ranked List with Scores

The top 10 books having maximum NS value are shown in bold in the table 3. Here the scores are coinciding; therefore, we are showing top 11 results. These top ranked products are shown in the table 4. The code P1, P2, etc. show the serial number of the products, the same product code is shown in both the table 3 and 4 respectively.

C. Final Ranked List of Books

The result, i.e. the top books for cloud computing are shown with the other books properties like Title, Author, Edition, Publisher, Price and Pages in table 5.

V. CONCLUSION

We intend to present a recommender technique to recommend top N products for some specific type, in this paper the product is book and N is 10.

We searched for top books on various topics and categorized the features; based upon these features we perform the ranking. The weights are given to several features and the weights are assigned according to the importance of the feature and requirement of the users.

Finally, top ten books are listed. Due to the limit of the page, only one topic, "cloud computing" is illustrated with results. These suggested ranked books are supposed to be one of the leading books on the topic concerned, and this work may helpful for millions of customers who seek for the best books available in the market.

In future, one can automate the proposed technique to reduce the workload and time.

REFERENCES

[1] http://www.internetworldstats.com/top20.htm

[2] A. Andreevskaia and S. Bergler,"Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses," EACL'06, pp. 209–216, 2006.

[3] G. Carenini, R. Ng, and A. Pauls, "Interactive Multimedia Summaries of Evaluative Text," IUI'06, 2006..

[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," KDD'04, 2004..

[5] N. Kaji and M. Kitsuregawa, "Automatic Construction of Polarity-Tagged Corpus from HTML Documents," COLING/ACL'06, 2006.

[6] A.M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," EMNLP-05, 2005.

[7] L. Zhuang, F. Jing, X.-Yan Zhu, and L. Zhang. Movie Review Mining and Summarization. CIKM-06, 2006..

[8] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1–2, 2008.

[9] V. Hatzivassiloglou and K. McKeown, "Predicting the semantic orientation of adjectives" EACL, pp- 174-181, 1997.

[10] S. Kim and E. Hovy,"Determining the sentiment of opinions," COLING, pp- 1367-1374, 2004.

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques" In EMNLP, pp- 79-86, 2002.

[12] Y. Fangy, L. Siy, N. Somasundaramy, Z. Yu, "Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model" WSDM'12,Seattle, Washingtion, USA, February 8–12, 2012.

[13] I. Ounis, M. Rijke, C. Macdonald, G. Mishne, and I. Soboro, "Overview of the TREC-2006 Blog Track," TREC, pp-15-27, 2006.

[14] Y. Seki, D. Evans, L. Ku, H. Chen, N. Kando, and C. Lin, " Overview of opinion analysis pilot task at NTCIR-6," NTCIR-6, pp-265-278, 2007.

[15] X. Ding, B. Liu, and P. Yu, "A holistic lexicon-based approach to opinion mining," WSDM, pp-231-240, 2008.

[16] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," WWW, pp-342-351, 2005.

[17] K. Dave, S. Lawrence, and D. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," WWW'03, 2003.

[18] M. Gamon, A. Aue, S. Corston-Oliver, and E. K. Ringger, "Pulse: Mining customer opinions from free text" IDA'2005.

[19] B. Pang and L. Lee, "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales," ACL'05, 2005.

[20] V. Hatzivassiloglou and J. Wiebe, "Effects of adjective orientation and gradability on sentence subjectivity," COLING, 2000.

[21] P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan, "An Exploration of Sentiment Summarization," Proc. of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2003.

[22] N. Kaji and M. Kitsuregawa, "Automatic Construction of Polarity-Tagged Corpus from HTML Documents," COLING/ACL'06, 2006.