ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

An Efficient Method for Determining Sentiment from Song Lyrics Based On WordNet Representation Using HMM

K.P Shanmugapriya, Dr.B.Srinivasan

Assistant Professor, PG & Research, Dept of Computer Science, Gobi Arts & Science College (Autonomous)

Gobichettipalayam, India

Associate Professor, PG & Research, Dept of Computer Science, Gobi Arts & Science College (Autonomous)

Gobichettipalayam, India

ABSTARCT: Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions. It plays major important role to analysis identify the emotions of the user. Earlier numbers of the works have been performed to analyze sentiments from speech, text and documents, from this the songs plays a most important to sentiment analysis since the songs and mood are mutually dependent to each other. Based on the selected song it becomes easy to find the mood of the listener, in future it will be used for recommendation systems. Songs are considered as a text file to perform a song it becomes imperative to find the hidden meaning of the song for mining the sentiment and classify them accordingly. Each song is a mixture of moods. In order to perform this process the input songs files are preprocessed using the semantic matching based on the WordNet Graph Representation and mining can be done by Hidden markov model (HMM) which classifies the topics into either two (positive/negative) or multiple (happy/angry/sad/...) classes. Topics mined by HMM can represent moods. For validation, we have used a dataset that consists of the different moods annotated by users of a particular website.

KEYWORDS: Music analysis, Sentiment mining, Variational inference, Similarity Measure, WordNet, Hidden Markov model (HMM).

I. INTRODUCTION

Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. The opinion expressed on the topic is given significance rather than the topic itself [1]. The main objective of Opinion mining is to determine the polarity of comments (positive, negative or neutral) by extracting features and components of the object that have been commented on in each document [2, 3]. Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic. . Sentiment mining is based on associating multiple sentiments with a document, which gives a more precise division into multiple classes, (*e.g.*, happy, sad, angry, disgusting, *etc.*) The work by Mihalcea, R [4], [5] in Sentiment Analysis is based on multiple sentiments. Data in the entertainment industry is mostly in the form of multimedia (songs, movies, videos, *etc*.). Listening to songs has a strong relation with the mood of the listener. A song is composed of both melody and lyrics. Both of them signify its emotions and subjectivity. Work has been done on melody as well as on lyrics for mining sentiments from songs [6]. A particular mood can drive us to select some song; and a song can invoke some sentiments in us, which can change our mood. Thus, song-selection and mood are interdependent features. Generally, songs are classified into genres, which do not reflect the sentiment behind them, and thus, cannot exactly estimate the mood of the listener. Thus there is a need to build an unsupervised system for mood estimation which can further help in recommending songs. Our goal is to use

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

HMM over song collections, so as to get topics, which probably correspond to moods, and give a sentiment structure to a song, which is a combination of different moods.

II. RELATED WORK

Lina Zhou et al., [8] investigated movie review mining using machine learning and semantic orientation. Supervised classification and text classification techniques are used in the proposed machine learning approach to classify the movie review. A corpus is formed to represent the data in the documents and all the classifiers are trained using this corpus. Thus, the proposed technique is more efficient. Though, the machine learning approach uses supervised learning, the proposed semantic orientation approach uses "unsupervised learning" because it does not require prior training in order to mine the data.

Jeonghee Yi et al., [9] proposed a Sentiment Analyzer to extract opinions about a subject from online data documents. Sentiment analyzer uses natural language processing techniques. The Sentiment analyzer finds out all the references on the subject and sentiment polarity of each reference is determined. The sentiment analysis conducted by the researchers utilized the sentiment lexicon and sentiment pattern database for extraction and association purposes.

Alekh Agarwal et al., [10] proposed a machine learning method incorporating linguistic knowledge gathered through synonymy graphs, for effective opinion classification. This approach shows the degree of influence among relationships of documents have on their sentiment analysis. This is brought about by the use of graph-cut technique and opinion words got through synonymy graphs of Wordnet.

Ahmed Abbasi et al., [11] proposed novel sentiment analysis methods to classify web forum opinions in multiple languages. The proposed sentiment analysis method utilized the function of stylistic and syntactic features to evaluate the sentiment in English and Arabic content. The Entropy weighted Genetic Algorithm is incorporated to enhance the performance of the classifier and achieve the true assessment of the key features. Experiments were conducted using movie review data set and the results demonstrated that the proposed techniques are efficient.

Bo Pang et al., [12] used machine learning techniques to investigate the effectiveness of classification of documents by overall sentiment. Experiments demonstrated that the machine learning techniques are better than human produced baseline for sentiment analysis on movie review data.

Qui et al., [13] analyzed the problems related to opinion mining such as opinion lexicon expansion and opinion target extraction. Opinion targets are entities and their attributes on which opinions have been expressed. The list of opinion words such as good, bad, excellent, poor used to indicate positive and negative sentiments is Opinion lexicon.

Gang Li & Fei Liu et.al [14] developed an approach based on the k-means clustering algorithm. The technique of TF-IDF (term frequency – inverse document frequency) weighting is applied on the raw data. Then, a voting mechanism is used to extract a more stable clustering result. The result is obtained based on multiple implementations of the clustering process. Finally, the term score is used to further enhance the clustering result. Documents are clustered into positive group and negative group.

Chaovalit and Zhou [15] compared the Semantic Orientation approach with the N-gram model machine learning approach by applying to movie reviews. They confirmed from the results that the machine learning approach is more accurate but requires a significant amount of time to train the model. In comparison, the semantic orientation approach is slightly less accurate but is more efficient to use in real-time applications.

II. PROPOSED METHODOLOGY

The major objective of the proposed system is to analyze the sentiment for mining attitude of each user for further recommendation. To perform this process collect the song lyrics, perform preprocessing steps such as tokenization,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

stemming, stop word removal, morphological ,average calculator and Word-Sense Disambiguation(WSD) using WordNet graphical representation and then find topics using Hidden Markov Model (HMM) thus each topic correspond to moods.

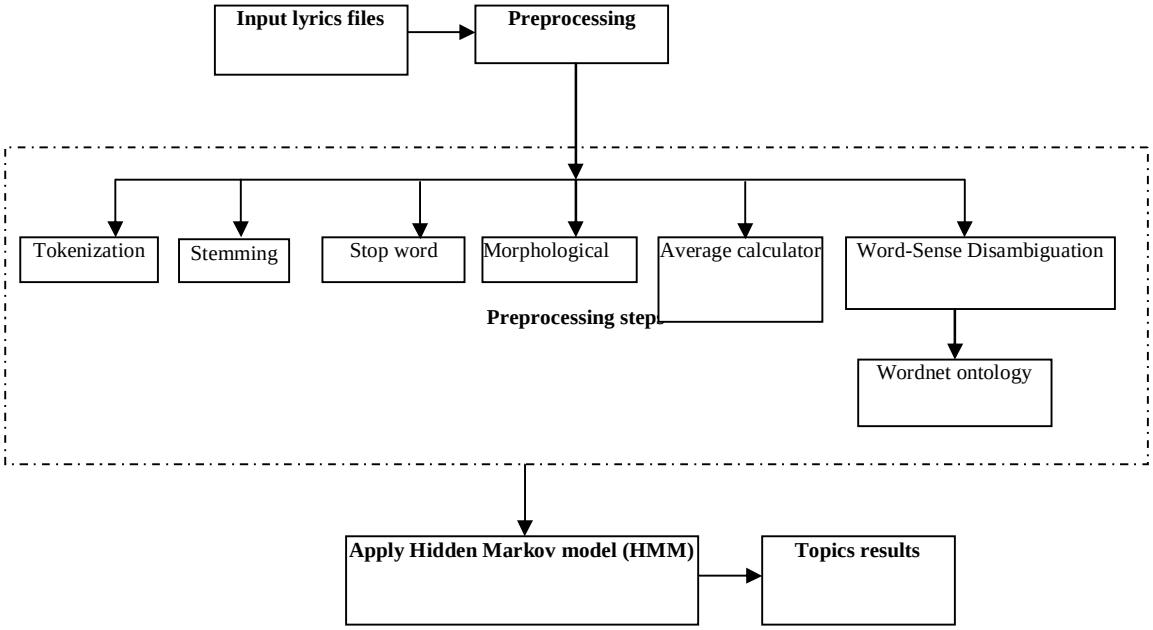

Figure 1: Block Diagram Representation of the Proposed System.

Tokenization is the first step in which the token from each and every songs of the text files were extracted by removing the white spaces, punctuation marks by calculating the occurrences of each token and it is stored in f_i.The Stop-word Removal is the Stop Words are words which do not contain important significance to be used in Search Queries. The Morphological Analysis provides the single word with different forms of the representation should be combined to form a single word. Since the word expresses the same meaning in every place. The Stemmer removes the diversity of words for indexing. Semantically meaningful word for the same words should be stemmed for data sparseness.

The Average Calculator calculates the average emotional words for each text file based on the moods of each song. The Word-Sense Disambiguation is the step in which word belongs to different meaning based on the situation is analyzed through their context [16]. To measure the semantic similarity of the each word with WordNet ontology [17], it assigns the semantic score value to each word. Furthermore, it returns the set of the synonyms for each word in order to increase the feature space [18].

The semantic relationship between the words is defined as , $d_{SIM}: w \times w \rightarrow \mathbb{R}^+$, where w be the set of the word presented in each song file existed in the graph G.The semantic relationship between the words for single lyrics text files is given by the relation, [19].

$$d_{SIM}(w_1, w_2) = \max_{w_1\#\in sens(w_1), w_2\#\in sens(w_2)} d_{SIM}(w_{1\#}, w_{2\#}) \quad (1)$$

Where $sens(w)$ describes the set of the words senses for each text files from song lyrics in the graph $w \in G, w_{\#}$ represents senses from the set of the senses associated for emotions. The finally measured semantic relationship of words is stored in WordNet index files, finally non-matching words presented in the text files are removed and uncommon words are also removed in this stage.

Automatic assessment of the contextual meaning of the words and the semantically measurement of the text objects [20] becomes a complex task. It is used to extend words with different meanings for analysis, and correlations between

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

the words can be measured directly. The general representation to solve the difficulty of contextual sense is specified through the set of key words from a phrase is represented in the form of,

$$P = \{w_1, \ldots w_n\} \quad (2)$$

Where,

P is the analyzed phrase,n is the total number of words presents in the text file f_1 ,w_n is the i word of the phrase P. The words presented in the phrase P , the set of senses is formed:

$$s_i = \{s_{i1}, \ldots s_{card(s_i)}\} \quad (3)$$

where $s_{ij}$is the j sense of the word. Let w_p ,be a polysemy word which measures the maximum semantic relationship among the word and another word found in the phrase for each text file ,it is represented as ,

$$w_p = \left\{ s_{pj} \middle| \begin{matrix} j = \arg \max \\ j = 1, Card(s_p) \end{matrix} \frac{\sum_{k \neq p, k=1}^{f} d_{SIM}(w_{pj}, w_k)}{f - 1} \right\} \quad (4)$$

Each polysemy word $w_p$is measured according to the words and senses founded in equation (3). The metric used for analysis each of WSD for each word presented in the file IC_{WSD} is defined using,

$$IC_{WSD} = \frac{\sum_{i=1}^{nr\_wsd} w_i}{nr\ wsd} \times 100 \quad (5)$$

nr_wsd is the number of polysemy words existing in the text file for each song and it is used for testing .w_i represents the relationship among the priory sense of the i word for each text file$F$, based on the formula

$$w_i = \begin{cases} 1 \text{ if } sens_{aprioric_i} = sensWSD_i \\ 0 \text{ Otherwise} \end{cases} \quad (6)$$

$sens_{aprioric_i}$ is the priory sense related to the word i for each text file F ,$sensWSD_i$ is generated by the WordNet ontology algorithm. The best semantic matching results found from WSD songs is represented as $SO = \{so_1, so_2, , , , , so_l\}$ which contains semantic meaningful words from WSD. For example, consider each entry of the song sf_i is represented a set of the ordered pair $(w_j, wsdn_{ji})$, signifying a word w_j and $wsdn_{ji}$ being the number of times that words semantically occurred in the WSD for song so_i.

The number of topics for each song is represented as k and applies the Hidden Markov Model (HMM) [21] to find topics for each song SO with words analyzed results from preprocessing steps; it is represented as B in the form of matrix $k \times v$ matrix. Each row in the matrix results from HMM denotes a different topic and each column represents a different word. HMM is used to estimate a set of songs in different states to find the find topic k of each song. In HMM the Markov process each one of the state is considered as number of words presents in the each song, lyrics text files f_i and observation state of the HMM is represented as classes which belongs to either positive, negative classes or multiple (happy/angry/sad/...) classes. The probability value of each topic K is calculated based on the multiple (happy/angry/sad/...) classes which are modeled in HMM. The usual definition of the HMM is as follows:

$$\lambda = (A, B, \pi) \quad (7)$$

Where A denotes the emission probability value for each word in the song without consideration of multiple (happy/angry/sad/...) classes, B denotes the transition probability results for each word to another with the consideration of multiple (happy/angry/sad/...) classes , π be the initial probability value for each word.

$$S = (s_1, \ldots s_N) \quad (8)$$

S be the number of words presents for songs results from preprocessing step, $V = \{happy, sad, angry\}$ is the set of the classes corresponds to the observed state, $Q = (q_1, \ldots . . q_T)$ and it is equal to the S with length T ,the observation state results of the classes is represented as $O = (o_1, \ldots o_T)$, transition probability matrix A following the state of the word j following state of the word i is represented as follows ,

$$A = [a_{ij}], a_{ij} = P(q_t = s_j | q_{t-1} = s_i) \quad (9)$$

B is the observation probability matrix for topic k result from state j multiple classes of time t,

$$B = [b_i(k)], b_i(k) = P(x_t = v_k | q_t = s_i) \quad (10)$$

π is the initial probability matrix

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

$$\pi = [\pi_i], \pi_i = P(q_1 = s_i) \quad (11)$$

To find topics for each song with multiple classes based on the following two steps:
The word of the each state is dependent on the previous state it is represented as follows:

$$P(q_t|q_1^{t-1}) = P(q_t|q_{t-1}) \quad (12)$$

The observation state results of multiple classes at time t is also dependent on the previous state, it is modeled as ,

$$P(o_t|o_1^{t-1}, q_1^t) = P(o_t|q_t) \quad (13)$$

The probability of the observations O for a specific state sequence Q with topic k is represented as follows:

$$P(O|Q,K) = \prod_{t=1}^{T} P(o_t|q_t,k) \quad (14)$$
$$= b_{q_1}(o_1) \times b_{q_2}(o_2) \dots b_{q_T}(o_T)$$

and the probability state sequence of the observation results of the different classes with topic k is

$$P(Q|k) = \pi_{q1,} a_{q1q2,} \dots . a_{qT-1qT} \quad (15)$$

$$P(O|k) \quad (16)$$
$$= \sum_{Q} P(O|Q,k)P(Q|k)$$
$$= \sum_{q1,\dots qT} \pi_{q1,} b_{q1,}(o_1) a_{q1q2} b_{q2}(o_2) .. a_{qT-1qT} b_{qT}(o_T)$$

This result allows the classification of the different topics based on the probability of O. In equation (13) the results of each topics result are observed by direct evaluation, some redundant probability results are occurring at this stage. In order to overcome this problem and reduce the time complexity of the HMM caching calculations is applied,

$$\alpha_t(i) = P(o_1, \dots . . o_t, q_t = s_i|k) \quad (17)$$

$\alpha_t(i)$ the sum of all observation probability results. This uses the following steps to calculate the values of the $\alpha_t(i)$ and find the topic k results in each song:
Initialization:

$$\alpha_1(i) = \pi_i b_i(o_1), 1 \le i \le N \quad (18)$$

Induction:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] \, b_j(o_{t+1}), 1 \le t \le T-1, 1 \le j \le N \quad (19)$$

Termination:

$$P(O|k) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] \, b_j(o_{t+1}), 1 \le t \le T-1, \quad 1 \le j \le N \quad (20)$$

For each state of the song s_j , $\alpha_j(t)$ stores the probability value of the each topic that observed at a time ,

$$P(O|k) = \sum_{i=1}^{N} \alpha_T(i) \quad (21)$$

To analysis the results of observation probability results for each topics k based on the backwards variable,

$$\beta_t(i) = P(o_{t+1} o_{t+2}, \dots . o_T|q_t = s_i, \lambda) \quad (22)$$

III. EXPERIMENTAL RESULTS

In order to perform the experimentation results collected the dataset from real time data that provided us with the lyrics for the songs. Assuming different values of k, ranging from 5 to 50, we applied HMM over the real time data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Precision

Precision is the percentage true positives results which are correctly classified for each songs.

$$Precision = \frac{tp}{tp + fp} \qquad (24)$$

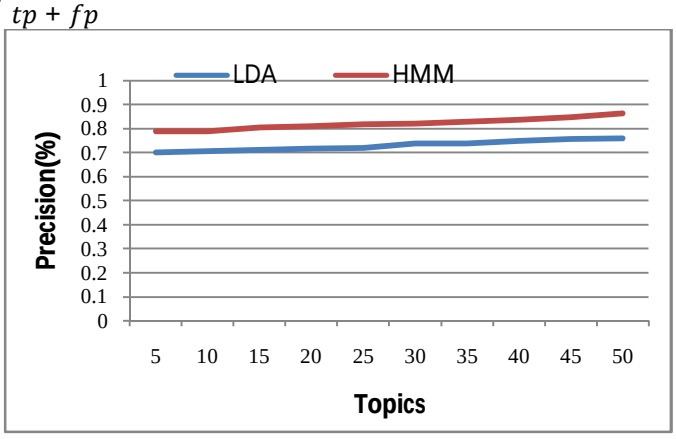

Figure 2: Precision measure the performance comparison results of precision between the proposed HMM with Wordnet based graph representation and existing LDA methods ,it can be seen that Precision value of the of the proposed HMM is high with the number of topics when compare to existing system ,it shows that proposed HMM have achieves higher mining results.

Recall

Recall in this context is also referred to as the true positive rate,

$$Recall = \frac{tp}{tp + fn} \qquad (25)$$

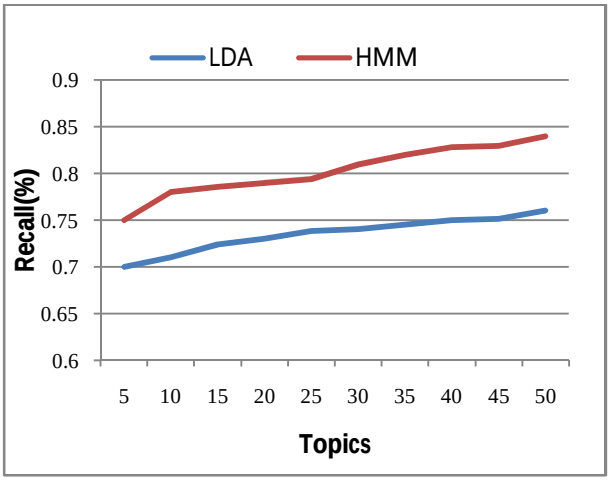

Figure 3: Recall measure the performance comparison results of recall between the proposed HMM with Wordnet based graph representation and existing LDA methods ,it can be seen that recall value of the of the proposed HMM is high with the number of topics when compare to existing system ,it shows that proposed HMM have achieves higher mining results

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

IV. CONCLUSION

The paper presents a sentiment analysis for mining the topics from songs based on their moods .The two words are co-occurred in the documents or input lyrics files are measure based on the wordnet graph representation. The sentiments of the each song are mined using Hidden Markov Model (HMM) .As far as songs are concerned, they reflect almost coherent sentiments and thus, form a good corpus for sentiment mining. Sentiments are a semantic part of documents, which are captured statistically by HMM, with positive results.

REFERENCES

1. Bing Liu," Exploring User Opinions in Recommender Systems", Proceeding of the second KDD workshop on Large Scale Recommender Systems and the Netflix Prize Competition", Aug 24, 2008, Las Vegas, Nevada, USA.
2. Dave.D, Lawrence.A, Pennock.D," Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", Proceedings of International World Wide Web Conference (WWW"03), 2003.
3. Turney, P, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", ACL"02, 2002.
4. Mihalcea, R.: A Corpus-based Approach to Finding Happiness. In: AAAI 2006Symposium on Computational Approaches to Analyzing Weblogs, pp. 139–144. AAAI Press, Menlo Park (2006)
5. Strapparava, C., Mihalcea, R.: Learning to Identify Emotions in Text. In: Proceedings of the 2008 ACM Symposium on Applied Computing, pp. 1556–1560. ACM, New York (2008)
6. Chu, W.R., Tsai, R.T., Wu, Y.S., Wu, H.H., Chen, H.Y., Hsu, J.Y.J.: LAMP, A Lyrics and Audio MandoPop Dataset for Music Mood Estimation: Dataset Compilation, System Construction, and Testing. In: Int. Conf. on Technologies and Applications of Artificial Intelligence, pp. 53–59 (2010)
7. Hsu, D.C.: iPlayr - an Emotion-aware Music Platform. (Master's thesis). National Taiwan University (2007).
8. [4] Lina Zhou, Pimwadee Chaovalit, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", Proceedings of the 38th Hawaii International Conference on system sciences, 2005.
9. [7] Yi, J., T. Nasukawa, R. Bunescu, and W. Niblack: 2003, "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques", in: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003). Melbourne, Florida.
10. [8] Alekh Agarwal and Pushpak Bhattacharyya, "Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified", In Proceedings of the International Conference on Natural Language Processing (ICON), 2005.
11. [9] Ahmed Abbasi, Hsinchun Chen, Arab Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums", ACM Trans. Inf. Syst., Vol. 26, No. 3. (June 2008), pp. 1-34.
12. [5] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.
13. Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen. "Opinion Word Expansion and Target Extraction through Double Propagation." Computational Linguistics, March 2011, Vol. 37, No. 1: 9.27.
14. Guang Qiu , Xiaofei He , Feng Zhang , Yuan Shi , Jiajun Bu , Chun Chen , "DASA: Dissatisfaction-oriented Advertising based on Sentiment analysis" , Expert Systems with Applications, 37 (2010) 6182–6191.
15. Chaovalit, Lina Zhou, Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches, Proceedings of the 38th Hawaii International Conference on System Sciences – 2005.
16. [31] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surv. vol. 34, pp. 1–47, 2002.
17. N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet," in Proc. ECAI'04, 2004, pp. 1089–1090.
18. C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, MA: Cambridge Univ. Press, 2008.
19. [10] Q. Peng, L. Zhao, Y. Yu, W. Fang, "A New Measure of Word Semantic Similarity based on WordNet Hierarchy and DAG Theory," International Conference on Web Information Systems and Mining, 2009, pg. 181-185, ISBN 978-0-7695-3817-4.
20. R. Richardson, A. Smeaton, J. Murphy, "Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Word", Technical Report, Working paper CA-1294, School of Computer Applications, Dublin City University, 1994.

BIOGRAPHY

K.P Shanmugapriya M.Sc., M.Phil., Assistant Professor, PG & Research Department of Computer Science, Gobi Arts & Science College (Autonomous), Gobichettipalayam – 638 453, Erode District, Tamil Nadu, India. She received her M.Phil Degree in Computer Science from Bharathidasan University in Junet-2004. She has authored or co-authored more than 8 conference presentations. Her research interests include Data Mining and Natural Language Processing.

Dr. B. SRINIVASAN M.C.A., M.Phil., M.B.A., Ph.D., Associate Professor, PG & Research Department of Computer Science, Gobi Arts & Science College(Autonomous), Gobichettipalayam – 638 453, Erode District, Tamil Nadu, India. He received his Ph.D. Degree in Computer Science from Vinayaka Missions University in 11.11.2010. He has authored or co-authored more than 70 technical papers and conference presentations. His research interests include automated biometrics, computer networking, Internet security, and performance evaluation.