ISSN: 2277-9655
Impact Factor: 4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

ONTOLOGY BASED ASPECT LEVEL OPINION MINING

Ms. Meleesa Alfonso*, Ms. Razia Sardinha
Department of Information Technology, Padre Conceicao College of Engineering , India
Department of Information Technology, Padre Conceicao College of Engineering , India

DOI: 10.5281/zenodo.56012

ABSTRACT

In recent years, opinion mining has been investigated mainly in three level of granularity (document, sentence or aspect(feature)). However both document and sentence level analysis do not discover what exactly customers liked or not. Due to very huge web size and growth rate, scalable and practical solutions are required. Studying opinion text, mainly aspect level is challenging.

In our project, a domain ontology has been introduced, which defines a space of hotel aspects, thus makes it possible for an hotel to be classified and scored by commonly accepted aspects. My approach thus enhances the user experience to search a hotel and compare it with other hotels aspect by aspect. The evaluation is based on the hotel reviews collected from traveler guide sites such as tripadvisor and makemytrip.

The basic idea of our approach is to capture the relationships among aspects, associations between aspects and their expressions of opinion. More specifically we utilize the domain ontology to construct a specific knowledge structure because it can clearly represent the certain relationships among domain concepts.

KEYWORDS: Aspect, Domain ontology, Ontology extraction, Opinion mining, sentiments.

INTRODUCTION

With the norm of consumers contributed data in the era of Web 2.0, increasingly more people have submitted or retrieved individual viewpoints about products, organizations via a variety of Web-based channels such as Blogs, forums, e-commerce sites and social networks. Due to problem of information overload [1], manually browsing a huge number of consumer reviews posted to the Web may not be achievable, if not totally impossible.

The massive volume of documents (e.g. customer reviews) archived on the Web has initiated the development of intelligent tools to automatically extract, examine and summarize their contents.

Opinion mining is also known as opinion analysis, sentiment analysis, or subjectivity analysis [2] [3]. Opinion analysis differs from Information Retrieval (IR) in that it aims at extracting the viewpoints about some entities rather than simply determine the topical information about those entities.

Analyzing the opinions or sentiments of consumer feedback posted to Blogs, forums, or e-Commerce sites can generate massive business values for organizations. Although consumer reviews are subjective in nature, these reviews are often well thought-out more valuable and trustworthy than other traditional information sources from the perspectives of customers.

In our project, we illustrate a novel opinion mining methodology which can automatically extract reviews, construct domain ontology, perform opinion mining and summarize consumers' reviews about various hotels with reference to the specific hotel contexts.

Though traditional opinion analysis was carried out at the document level, increasingly more research has examined sentiment analysis at the more fine-grained sentence level in recent years. Even if a review (i.e., document) is rated as positive, negative opinion words could appear in the same review. Therefore, opinion mining against consumers' reviews is often performed at the aspect level to provide deep analytics for the target entity [6] [7] [8]. The request

for a fine-grained opinion mining method is driven by the fact that sentiment words are often context-dependent. Ontology is generally considered as a formal specification of conceptualization which consists of concepts and their relationships[20]. Domain ontology is one kind of ontology which is used to represent the knowledge for a specific type of application domain (e.g., a hotel domain). Our model of fuzzy domain ontology is underpinned by fuzzy sets and fuzzy relations [21] as defined below:

Definition 1. **Fuzzy Set:** A fuzzy set F consists of a set of objects drawn from a domain M and the membership of each object m_i in F is defined by a membership function

$\mu_f : M \rightarrow [0,1]$.

Definition 2. Fuzzy Relation: A fuzzy relation R_{XY} is defined as the fuzzy set R on a domain M ×N where M and N are two crisp sets. The membership of each object (m_i, n_i) in R is defined by a membership function $\mu_f : M \times N \rightarrow [0,1]$.

Definition 3. **Fuzzy Domain Ontology**: A fuzzy domain ontology is a triple Ont = (C, R_{NTAX}, R_{TAX}) where C is a set of concepts (classes). The fuzzy relation $R_{NTAX} : C \times C \rightarrow [0,1]$ defines the strength of the non-taxonomic relationship for each pair (c_i, c_i) in R_{NTAX}, and the fuzzy relation $R_{TAX} : C \times C \rightarrow [0,1]$, defines the strength of the taxonomic (subclass/super-class) relationship for each pair (c_i, c_i).

With reference to our application, C represents the set of hotels, hotel aspects, sentiments, and so on.

Ontology is often specified in a declarative form by using semantic markup languages such as RDF and OWL [22]. Ontology provides many potential benefits in representing and processing knowledge, including the separation of domain knowledge from application knowledge, sharing of common knowledge of subjects among humans and computers, and the reuse of domain knowledge for a variety of applications.

Linguistic or inference based methods can deal with sentiment analysis for some general cases, but there are many instances (particularly down to the phrase level) that the general rules or inference process could not be applied. For example, no general linguistic rule can be applied to detect the polarity of the sentiment 'small' in the sentence 'The hotel is good in general; the rooms are small'. On the other hand, machine learning methods usually require a huge number of manually labelled training examples to build an accurate classifier. Nevertheless, manually annotating a huge number of review messages at the sentence level is extremely labour intensive and expensive. Although attempts are made to mine consumers' reviews at the aspect level, the polarities of sentiments are assumed the same across product domains (i.e., context-free). For instance, 'small' is often assumed negative no matter it is referring to a hotel room or the size of a Netbook computer. Indeed, it has been pointed out that developing an automatic technique for building opinion lexicon is an important topic for research and practices in opinion mining, and contextual domain knowledge is important to improvise the performance of opinion mining systems.

The main contributions of our research are:

(1) The design of a novel Fuzzy domain ontology consisting of concepts and attributes associated with the concepts and the taxonomic and non-taxonomic relationships between them.

(2) Using the domain ontology during the aspect selection stage in opinion mining and extracting the sentiments associated with the aspects.

(3) Scoring the sentiments associated with the aspects to get the total score for the entity of interest.

(4) Comparing different entities aspect by aspect.

RELATED WORK

The traditional research of opinion mining (or sentiment analysis) is defined as the task of the sentiment classification at document-level. However, for many opinion expressions such as twitter, micro blogs, and customer feedback reviews only judging the sentiment orientation is not enough. Therefore, increasingly more research has examined opinion mining at the sentence, phrase level and more fine-grained aspect or feature-level in recent years. Aspect-level opinion mining (also called feature or aspect-level sentiment analysis) is the research problem that focuses on the recognition of all sentiment phrases within a document (e.g. customer review) and the aspects to

which they refer. The hotel reviews which people commented have many aspects (features) and different opinions about each aspect.

A light weight fuzzy domain ontology extraction method has been developed to automatically create concept hierarchies based on textual contents extracted from online reviews.

The algorithm of fuzzy domain ontology extraction includes concept extraction, concept pruning, dimensionality reduction, and fuzzy relation extraction. Fuzzy relation extraction involves the generation of taxonomic relations using the structural similarity (SSIM) metric developed in the field of image analysis. Formal concept analysis[14] and fuzzy formal concept analysis have also been applied to build domain ontology automatically. Formal concept analysis[4] is a systematic method for deriving implicit relationships among concepts described by a set of attributes. For the research work reported in this paper, we utilize a simplified version of the fuzzy domain ontology model for sentiment knowledge representation. In particular, we develop effective computational methods to learn the non-taxonomic relations among concepts (e.g., hotel, hotel aspects and sentiments) to support opinion mining.

An econometric opinion mining method has been proposed to analyze product aspect evaluations expressed in online consumer reviews[6]. Each product aspect is represented by a noun which frequently appears in the users reviews. A manual procedure is then involved to filter the candidate nouns to identify correct product aspects. The adjectives collocated with product aspects are taken as the sentiment words. A pair of product aspect and sentiment (also called an opinion phrase) is formally represented by a vector in the tensor product space. Hedonic regressions are applied to estimate the relative weights of product features and the strength of the sentiments associated with those features. OPINE employs the 'relaxation labeling' classification method developed by the computer visioning research community to detect sentiment polarity[8]. Similarly, Feature-Based Summarization (FBS) system has been developed to extract explicit product aspects and sentiments at the sentence level [7]. The Apriori association rule mining algorithm is applied to extract the product aspects (i.e., noun phrases) frequently occurring in product reviews. A similar product aspect extraction method is also applied to a product review mining system [9]. The ReviewSeer system adopts an n-gram approach for aspect extraction and a machine learning approach for sentiment polarity classification[15]. For the aforementioned opinion mining systems, polarity detection of sentiments is not carried out with respect to a particular product domain.

Entropy Weighted Genetic Algorithm (EWGA) has been developed to select the best syntactic (e.g., POS pattern) and stylistic features (e.g., number of special characters used in a document) for multilingual (e.g., English and Arabic) sentiment classification against various extremist online forums[2]. The EWGA algorithm selects the most informative features (e.g., n-gram1) according to information gain and passing those features to a SVM classifier for polarity classification (e.g., positive or negative) at the document level. Based on the technique of bootstrapping, a classification accuracy of 91% is achieved over a benchmark movie dataset[16].

In the field of IR, Probabilistic Latent Semantic Analysis (PLSA) which is underpinned by the unigram language modeling approach is proposed to predict sentiment orientations in movie blog posts[13] . The PLSA model is combined with a time series analysis model (called autoregressive model) to predict the gross revenues of movies. PLSA is also applied to combine opinions expressed in a well-written expert review with those retrieved from Web 2.0 sources such as blog posts to generate a comprehensive opinion summary about a product or a political figure. Probabilistic generation language models are explored to identify and rank sentiment expressions at the document level.

We propose to address the problem of opinion mining using a fuzzy approach e.g., modeling the association between a hotel feature and a sentiment in terms of a fuzzy relation. In the field of machine learning, the problem of automatically identifying sentiment orientations across different domains is called the 'Domain-Transfer' problem[17]. A method called Relative Similarity Ranking (RSR) is proposed to select the most informative unlabeled opinionated documents from a training set to re-train a classifier (e.g., Support Vector Machine). Instead of identifying the most informative training examples, We employ the available sentiment lexicons such as SentiWordNet.

Linguistic rules are applied to detect the context-sensitive orientations of sentiments or opinions extracted from online customer reviews[18]. For example, for the sentence 'The camera takes great pictures and has a long battery life', the orientation of the sentiment 'long' is classified as positive because it is associated with the positive seeding sentiment great. An inference-based opinion mining method called Semantic Orientation (SO) analysis has been developed to compute the polarity of sentiments[19]. The SO of an arbitrary word can be estimated based on the strength of association between the word and fourteen seeding sentiment words such as good, nice, bad, poor, and so on. Point-wise Mutual Information (PMI) is proposed to compute the strength of association between any pair of words.

Our system also employs a variant of Mutual Information to estimate the strength of associations between product features and sentiment words.

Context-sensitive sentiment analysis has been an active research topic in the Natural Language Processing (NLP) community[23] . A sentence is first parsed and represented by a dependency tree. A set of linguistic features are used to train the AdaBoost classifier to predict the sentiment orientation of a target word. An appraisal group is represented by a set of attribute values in some task-independent semantic taxonomies such as attitude, orientation, graduation, and polarity[24]. The appraisal group method has been applied to analyze the sentiments of a movie review corpus.

Apart from utilizing the fuzzy domain ontology, our system also employs basic sentiment lexicons to infer sentiment polarity. However, instead of using sophisticated NLP techniques which are computationally expensive, we adopt a light-weight NLP approach so that my opinion mining system can scale up for the sheer volume of customers contributed feedback data generated in the era of Web 2.0.

ONTOLOGY BASED ASPECT LEVEL OPINION MINING

The general system architecture of my Ontology-Based Aspect Level Opinion Mining System is shown in figure 1.

Figure:

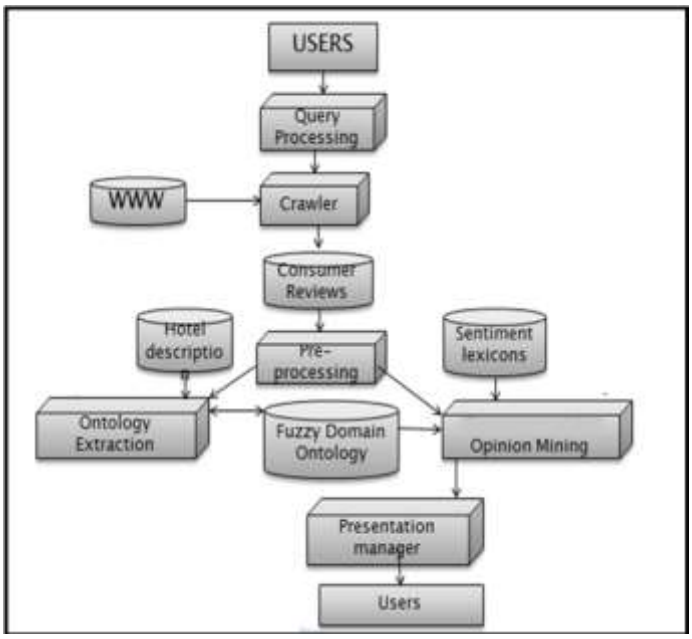


Fig. 1 Overall System architecture

ISSN: 2277-9655
Impact Factor: 4.116

Pre-processing

The user first selects a specific hotel for opinion mining. Based on the selected hotel, the crawlers can be invoked to download consumer reviews and hotel descriptions related to that specific hotel. Document pre-processing techniques such as stopword removal and POS tagging are then invoked to process the consumer reviews and hotel descriptions.

Fuzzy Domain Ontology Extraction

The Fuzzy Domain Ontology extraction is carried out offline and must be performed before opinion mining is conducted. It captures taxonomic information and non-taxonomic relationships. Consumer reviews and hotel descriptions are fed into this ontology extraction module. The standard document pre-processing techniques are applied to each hotel review and description documents. Then a windowing process is carried out over the collection of documents. The windowing process helps in reducing noisy term relationships.

For each document a virtual window of δ words is moved from left to right one word at a time until the end of a sentence is reached. The statistical information among tokens within each window is collected to construct collocational expressions. This process is repeated for each document until the entire collection has been processed. Only the specific linguistic pattern (Adjective Noun and Noun Noun) defined are analysed. If a token has an association weight less than pre-defined threshold value, it will be rejected.

Balanced Mutual Information(BMI) method is used to compute the degree of association among tokens. This method takes into account both term presence and term absence as an evidence of the implicit term relationships.

Formulae:

$$\mu ci(t_n) \approx BMI(t_m, t_n)$$
$$= \beta(P_r(t_m, t_n) \log2(P_r(t_m,t_n)/P_r(t_m)P_r(t_n)))$$
$$+P_r(\neg t_m,\neg t_n) \log2(P_r(\neg t_m,\neg t_n)/P_r(\neg t_m)P_r(\neg t_n)) - (1-\beta)(P_r(t_m,\neg t_n) \log2(P_r(t_m,\neg t_n)/P_r(t_m)P_r(\neg t_n)))$$
$$+P_r(\neg t_m, t_n) \log2(P_r(\neg t_m,t_n)/P_r(\neg t_m)P_r(t_n))) \quad (1)$$

where $\mu ci(t_n)$ is the membership function to estimate the degree of a term $t_n \in X$ belonging to a concept $ci \in C$. $\mu ci(t_n)$ is the computational mechanism for the the relation RXC defined in the fuzzy ontology Ont =<X, C, RXC , RCC >. The membership function $\mu ci(t_n)$ is indeed approximated by the BMI score. $P_r(t_m, t_n)$ is the joint probability that both terms appear in a text window, and $P_r(\neg t_m,\neg t_n)$ is the joint probability that both terms are absent in a text window. Terms in the potential concept with membership less than the threshold are discarded.

The relevance score for the concepts is calculated and only the concepts with relevant score greater than the threshold are retained. For each selected concept, its context vector will be expanded based on the synonymy relation defined in WordNet. Finally the fuzzy taxonomy is generated based on the subsumption relations among extracted concepts.

Opinion Mining

The opinion mining module uses the fuzzy domain ontology and sentiment lexicons to extract the most frequent aspects corresponding to the aspects in the ontology. We first extract the aspects from the pre-processed reviews. We than load the domain ontology and get the sentiments and the sentiment scores(positive or negative) associated with only those aspects present in the domain ontology. Thus each aspect gets a score.

Summing up the scores of each aspect we get the total score for the hotel. The results are the presented to the users.

RESULTS AND IMPLEMENTATION

We first construct the fuzzy domain ontology using the various hotel descriptions and reviews(Fig. 2). We than select a hotel to perform opinion mining. We extract customer reviews of the hotel and store each review as a flat text file. We than perform pre-processing techniques on each review.The fuzzy domain ontology is than loaded and we extract only the hotel aspects from the reviews present in the ontology. The sentiments and their scores

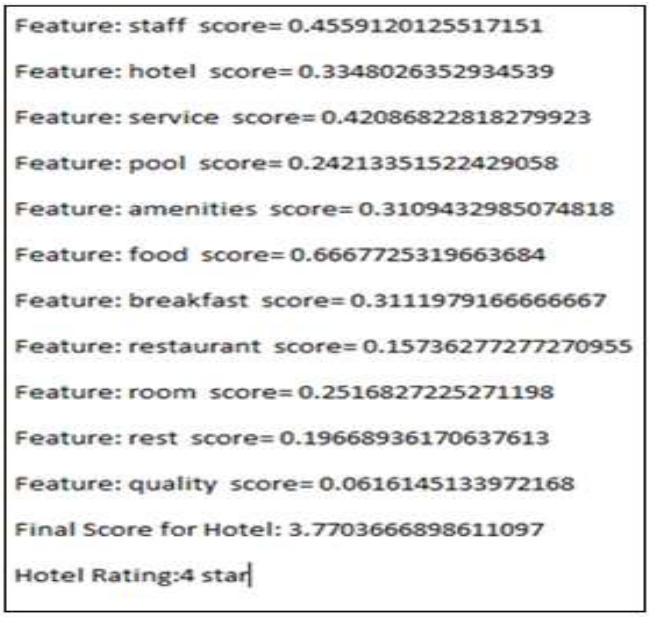
associated with each aspect are extracted. Summing the sentiment scores associated with each aspect we get total score for each aspect. Finally summing up the scores of each aspect we get the score for the hotel. Based on the score we give a rating to the hotel(Fig. 3).

Figure:

```
<!-- http://www.semanticweb.org/ontologies/concetti#hotel -->

<owl:Class
rdf:about="http://www.semanticweb.org/ontologies/concetti#hotel">
<rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/ontologies/concetti#amen
ities"/>
<rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/ontologies/concetti#fant
astic"/>
<rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/ontologies/concetti#hote
l"/>
<rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/ontologies/concetti#pool
"/>
<rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/ontologies/concetti#rest
aurant"/>
<rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/ontologies/concetti#staf
f"/>
</owl:Class>
```

Fig. 2 Snapshot of hotel ontology



```
Feature: staff score= 0.4559120125517151
Feature: hotel score= 0.3348026352934539
Feature: service score= 0.42086822818279923
Feature: pool score= 0.24213351522429058
Feature: amenities score= 0.3109432985074818
Feature: food score= 0.6667725319663684
Feature: breakfast score= 0.3111979166666667
Feature: restaurant score= 0.15736277277270955
Feature: room score= 0.2516827225271198
Feature: rest score= 0.19668936170637613
Feature: quality score= 0.0616145133972168
Final Score for Hotel: 3.7703666898611097
Hotel Rating:4 star
```

Fig. 3 Result of opinion mining for a hotel

Similarly we can perform opinion mining on other hotels and compare them aspect by aspect.

ISSN: 2277-9655
Impact Factor: 4.116

CONCLUSION

The popularity of the social web has had a tremendous impact on a number of different research topics. Particularly, the possibility of extracting numerous kinds of value-added, informational elements from customers' opinions has attracted researchers from various domains like information retrieval and computational linguistics. This process is called opinion mining or sentiment analysis and is the most challenging research field in this area.

In our project, we propose an innovative methodology for aspect-based opinion mining using a fuzzy ontology. The proposed approach is based on three different stages: (i) construction of a fuzzy domain ontology(hotel); (ii) using the ontology for aspect identification; (iii) assigning polarity to each aspect based on SentiWordNet.

The ontology is used to improve the process of aspect identification. The ontology contains concepts, attributes and relationships. In due course of this research, Data Mining and Natural Language Processing techniques relevant to each phase of the project were studied. First the reviews were crawled and stored separately in a text file. Each text file was pre-processed and stored in a structured format. These structured reviews were then used to extract only those Hotel aspects present in the ontology. In the next stage of this research, the Sentiments adjacent to the aspects were extracted. A summary of sentiment scores was generated for each aspect and using these scores, a total score for the hotel was calculated. Based on the score the hotel was given a star rating.

ACKNOWLEDGEMENTS

The success of this report depends largely on the encouragement and guidelines of many people. This research would not have been possible if I had no support of many individuals and organization. I therefore would like to extend my sincere gratitude to all of them who made it possible for me to complete my project successfully.

I express my heartfelt gratefulness to Mrs. Razia de Loyola Furtado e Sardinha, Assistant Professor in IT Engineering, for her stimulating supervision, continuous guidance, suggestions whenever required during my dissertation work. I am thankful to her for spending her valuable time reviewing my project work and encouraging me at every step of my project.

I thank my parents and friends for their support and encouragement throughout my project work. I am grateful to God for his blessings for which words are not enough to praise his glory.

REFERENCES

[1] Lau, R.Y.K., Bruza, P.D., and Song, D. "Towards a Belief Revision Based Adaptive and Context Sensitive Information Retrieval System", ACM Transactions on Information Systems, (26:2), March 2008, pp. 8:1-8:38.

[2] Abbasi, A., Chen, H., and Salem, A. "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums", ACM Transactions on Information Systems (26:3), June 2008, Article 12.

[3] Turney, P. and Littman, M. "Measuring praise and criticism: Inference of semantic orientation from association", ACM Transactions on Information Systems, (21:4), October 2003.

[4] Tho, Q.T., Hui, S.C., Fong, A., and Cao, T.H. "Automatic Fuzzy Ontology Generation for Semantic Web", IEEE Transactions on Knowledge and Data Engineering, (18:6), June 2006.

[5] Lau, R.Y.K., Song, D., Li, Y., Cheung, C.H., Hao, J.X. "Towards A Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning", IEEE Transactions on Knowledge and Data Engineering, (21:6), 2009.

[6] Hu, M. and Liu, B. "Mining and summarizing customer reviews", in Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, 22-25 August 2004.

[7] Popescu, A.M. and Etzioni, O. "Extracting Product Features and Opinions from Reviews", in Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada, October 2005.

[8] Miao, Q., Li, Q., and Dai, R. "An integration strategy for mining product features and opinions", in Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, October 2008.

[9] "A Fuzzy Domain Sentiment Ontology based Opinion Mining Approach for Chinese Online Product Reviews"Hanshi Wang, XinhuiNie, Lizhen Liu College of Information Engineering, Capital Normal University, Beijing, China Lizliu12409.
[10] "A Peer Review of Feature Based Opinion Mining and Summarization"Padmapani P. Tribhuvan, S.G. Bhirud ,Amrapali P. Tribhuvan.
[11] "Domain Ontology Construction by Partial Import for Document Annotation"TayybahKiren, Muhammad Shoaib, and Sang-Jo Lee.
[12] Lau, Raymond Y.K. and Li, Yuefeng and Xu, Yue (2007) "Mining Fuzzy Domain Ontology from Textual Databases". In Proceedings IEEE/WIC/ACM International Conference on Web Intelligence, pages pp. 156-162, Silicon Valley, USA.
[13] Liu, Y., Huang, X., An, A., and Yu, X. "ARSA: A Sentiment- Aware Model for Predicting Sales Performance Using Blogs", in Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, Netherlands, 23-27 July 2007, pp. 607-614.
[14] Cimiano, P., Hotho, A., and Staab, S. "Learning concept hierarchies from text corpora using formal concept analysis", Journal of Artificial Intelligence Research (24), 2005, pp. 305–339.
[15] Dave, K., Lawrence, S., and Pennock, D. "Mining the peanut gallery: opinion extraction and semantic classification of product reviews", in Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, 20-24 May 2003, pp.519-528.
[16] Pang, B., Lee, L., and Vaithyanathan, S. "Thumbs up? Sentiment Classification using Machine LearningTechniques", in Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing,Philadelphia, 2002, pp. 79-86.
[17] Tan, S., Wang, Y., and Cheng, X. "Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples", in Proceedings of the 31st annual international ACM SIGIR AnnualInternational Conference on Research and Development in Information Retrieval, Singapore, 20-24 July 2008,pp.743-744.
[18] Ding, X. and Liu, B. "The Utility of Linguistic Rules in Opinion Mining", in Proceedings of the 30th AnnualInternational ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam,Netherlands, 23-27 July 2007, pp. 811 - 812.
[19] Hatzivassiloglou, V. and McKeown, K. "Predicting the semantic orientation of adjectives", in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 1997, pp. 174 - 181.
[20] T. R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199– 220, 1993.
[21] Zadeh, L.A.. "Fuzzy sets", Journal of Information and Control, (8), 1965, pp. 338-353.
[22] The World Wide Web Consortium. Web Ontology Language, 2004. Available from http://www.w3.org/2004/OWL/.
[23] Wilson, T., Wiebe, J., and Hwa, R. "Recognizing strong and weak opinion clauses", Computational Intelligence(22:2), 2006, pp. 73-99.
[24] Whitelaw, C., Garg, N. and Argamon, S. "Using appraisal groups for sentiment analysis", in Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, October 2005, pp. 625 - 631.