© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Survey : Ontology Based Information Retrieval For Sentiment Analysis

Gopi A. Patel , Nidhi Madia

Computer Engineering Department, Silver Oak College of Engineering and Technology, Ahmedabad, Gujarat, India

ABSTRACT

The rapidly growing data on the web has created a big challenge for directing the user to the web pages in their areas of interest. Sentiment analysis or Opinion mining plays an important role in finding the area of interest based on user's previous actions. Social networking portals have been widely used for expressing opinions in the public domain. Text based sentiment classifiers often prove inefficient. Semantic web is the solution for Searching relevant information from huge repository of unstructured web data. Semantic web leads the idea of ontology as background knowledge represents the concepts and the relationship in specialized domain. The basic idea behind this survey is to take domain ontology for providing more elaborate sentiment scores. We discuss an approach where information retrieved from web and ontology is created before sentiment classification and focuses on how to classify the semantic orientation of text.

Keywords: Ontology, Sentiment Analysis, Semantic Web, Web Mining

I. INTRODUCTION

The web contains huge collection of unstructured data which makes difficult to retrieve the relevant information. Nowadays number of internet users has grown considerably over the past decade and continues to increase. Along with the number of users, data available on the internet continues to increase exponentially. An important part of our information-gathering behavior has always been to find out what other people think. So mining the user opinion and sentiments are very useful in many applications.

Sentiment analysis is one of the key emerging technologies in the effort to help people navigate the huge amount of user generated content available online [1]. Social Networks have become one of the attractive communication medium used over internet. Millions of text messages are appearing daily on popular web-sites that provide micro-blogging services such as Facebook, Twitter. Consequently, micro-blogging web sites have become rich data sources for opinion mining and sentiment analysis. Text-based sentiment classifiers often prove inefficient. So ontology based techniques can be used for sentiment analysis. Ontologies as background knowledge can be used to improve the process and results of Web Mining for finding user behavioral patterns on World Wide Web [7]. In this approach ontology is used to present domain knowledge about the subject of opinion. It allows showing the structure of product or servicing which is rated in opinion. In another ontology-based approach to sentiment analysis the ontology of sentiment can be created [9].Many algorithms and methods are used to create ontology for sentiment analysis. In this paper we discuss an approach where information retrieved from web and ontology is created before sentiment classification and focuses on how to classify the semantic orientation of text. In the next sections, we give brief summary of the areas Web Mining, Ontology, Semantic Web, and Sentiment Analysis.

II. METHODS AND MATERIAL

2. Semantic Web Mining

2.1 Semantic Web:

Semantic Web Mining combines Semantic Web and Web Mining. The Semantic Web is based on a vision of Tim Berners-Lee, the inventor of the WWW [3]. The great success of the current WWW leads to a new

challenge that a huge amount of data is interpretable by humans only; machine support is restricted. Berners-Lee suggests enriching the Web by machine-processable information which supports the user in his tasks. For instance, today's search engines are already absolutely powerful, but still too often return very large or inadequate lists of hits. Machine-processable information can point the search engine to the relevant pages and can hence improve both precision and recall. Semantics as a word means the "study of meanings" [2]. Berners-Lee suggested a layer structure for the Semantic Web [3].

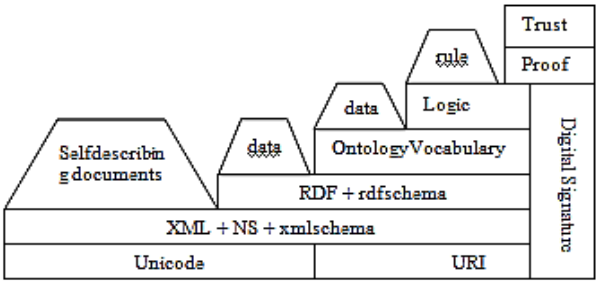

Figure 1. The layers of Semantic Web

Unicode and URl: The concept of Unicode is the standard used for the presentation of computer character. This provide a starting point for representing characters used in most of the languages in the world, URls on the other hand provide the standards for identifying and locating resources such as pages on the Web.

XML**:** XML and its related standards such as Schemas and Namespaces form a common medium for structuring data on the Web. They do not communicate the meaning of the data, but nevertheless are well established within the Web [4].

Resource Description Framework**:** RDF is the first layer of the Semantic Web. RDF is a simple metadata representation framework. It can be seen as the first layer where information converted into machine-understandable. RDF "is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web" [3].

RDFSchema: RDF Schema is a simple type modeling language that used for describing the classes of resources along with their properties in the basic RDF model. It delivers a simple reasoning framework for inferring the types of resources.

Ontologies: Ontology is a richer language that providing more complex constraints on the types of resources and their properties. Ontology is an explicit formalization of a shared understanding of a conceptualization [5].

Logic and Proof**:** An automatic system provided on top of the ontology structure to make new inferences. Thus, using such a system, a software agent can make assumption as to whether a particular resource satisfies its requirements or not (and vice versa).

Logic and Proof**:** An automatic system provided on top of the ontology structure to make new inferences. Thus, using such a system, a software agent can make assumption as to whether a particular resource satisfies its requirements or not (and vice versa).

Trust**:** The motive of last layer of the layered architecture is to know trustworthiness of the information by asking questions in Semantic Web. This gives the quality assurance of that information. The Semantic Web uses in reasoning while searching towards the precision of data on web for the search query so we can say that Semantic Web helps the Web machine process better.

2.2 Web Mining

Web mining consists of a set operations defined on data residing on WWW data servers. Mobasher et al. defines web mining as "the discovery and analysis of useful information from the World Wide Web" [10]. The constant growth in the size and the use of the World Wide Web presented new techniques for processing these large amounts of data. The incentive of web mining is to mine the data available on the internet is quite strong. In web mining all data mining techniques are applied on web. Web mining is mainly categorized into three subsets [10]:

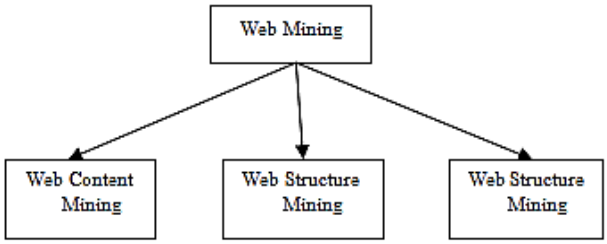

Figure 2. Types of Web Mining

1. Web content mining: Web content mining used to analyze the content of Web resources. It focuses on the discovery of patterns in large document collections [3]. The focus is on the content of web pages themselves. This includes all types of data text, images, audio file, and etc. Webmining extracts information from content of the web document.
2. Web Structure Mining: Web structure mining focuses on the links rather than the content of the pages, their usage or semantics. The hyperlinks that link the web pages and the document structure itself such as the xml or html structure. So it aims to analyze the process in which different web documents are linked together.
3. Web Usage Mining: Usage mining as the name implies focus on how the users are interact with web site, the web pages visited, the order of visit, timestamps of visits and time durations of them. The main source of data for the web usage mining is the server logs which log each visit to each web page with possibly IP, referrer, time, browser and accessed page link [6]. It is also known as log mining, because it includes mining the web server logs.

2.3 Semantic Web Mining

Semantic web mining is a combination of the two fast-growing research areas Semantic Web and Web Mining .Both are built on the success of the World Wide Web. They complement each other well because they each address one part of a new challenge posed by the great success of the current WWW [6]. Large amount of the data on the Web is so unstructured that they can only be understood by humans, but the amount of data is so large that they can only be processed efficiently by machines. The Semantic Web addresses the first part of this challenge by trying to make the data machine-understandable, while Web Mining addresses the second part by automatically extracting the hidden useful knowledge in these data, and making it available as an aggregation of manageable proportions. This view follows our observation that trends converge in both areas that increasing numbers of researchers work on improving the results of Web Mining by utilizing semantic structures in the Web, and make use of Web Mining techniques for building the Semantic Web [3]. It can be read both as Semantic (Web Mining) and as (Semantic Web) Mining.

2.4 Ontology

Ontology is a Greek word meaning study/science (logy) of being (onto) [6]. The word "ontology" has been recognized in philosophy as the subject of existence. It is a sub-branch of philosophy. Ontology is "an explicit formalization of a shared understanding of a conceptualization" [3]. Ontology concepts and the relationship between those concepts should be explicitly defined. Ontology should be machine-readable and the ontology should capture related knowledge accepted by the community. Ontology will play an important role in the second generation of the web, which Tim Berners-Lee call the "Semantic Web". Information in semantic web is given well-defined meaning, and is machine-readable. Search engines will use ontology to find pages with words that are syntactically different but semantically similar [7]. Ontology is "A specification of a representational vocabulary for a shared domain of discourse — definitions of classes, relations, functions, and other objects — is called ontology" [6]. Ontology can also defines multiple relations between entities, restrictions, classes and also the way these relations are to be used.

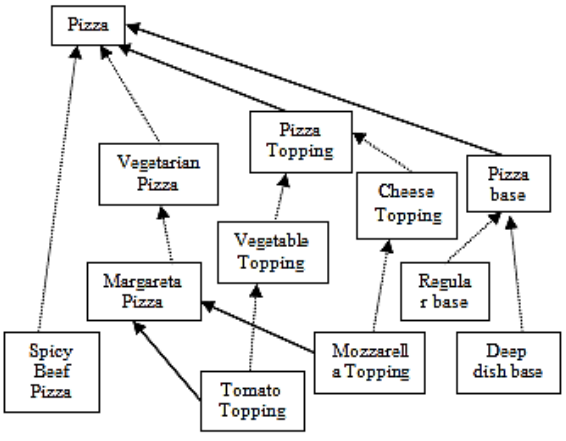

Figure 3. Example of Ontology

III. RESULTS AND DISCUSSION

3. **Sentiment Analysis**

Sentiment analysis is also known as opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions regarding entities such as products, services, individuals, issues, events, topics, organizations, and their attributes. Sentiment analysis or opinion mining mainly focuses on opinions which express positive or negative sentiments. It refers to the application of natural language Processing, computational linguistics, and text analytics to identify and extract subjective information in source material [8].The field of sentiment has become a very active research area. There are many reasons for this. First, it has a wide arrange of applications, almost in every domain. Second, it offers many challenging research problems, which had never been studied before. Whenever we need to make a decision on some specific topic, we want to know others' opinions. In the real world, businesses and organizations always want to find the opinions of consumers or public about their products and services. Opinions are key influencers of our behaviors so they central to almost all human activities.

Sentiment classification is generally composed as a two-class classification problem, positive and negative. Training and testing data used are normally product reviews. Online reviews have rating scores assigned by their reviewers, e.g., 1-5 stars, the positive and negative review classes are determined using the ratings. For example, a review with 4 or 5 stars is considered a positive review and a review with 1 or 2 stars is considered as a negative review. Sentiment classification is essentially a text classification problem. In this sentiment or opinion words that indicate positive or negative opinions are more important such as great, excellent, amazing, horrible, bad, worst, etc. There are various machine learning approaches described below.

3.1 Naïve Bayes

The Naive Bayes classifier is a probabilistic model based on the Bayes' theorem, which calculates the possibility of a tweet belonging to a specific class such as neutral, positive or negative [15].

3.2 Random Forest

It is also known as tree classifiers, which are used to predict the class based on the categorical dependent variable. This classifier's error rate depends on the correlation between any two trees in the forest and the strength of the each individual tree in the forest.

3.3 Support Vector Machines (SVM)

SVMs are the class of algorithms that are based on kernel substitution. This is a type of the supervised learning algorithm. This will be trained with a learning algorithm that implements a learning bias derived from statistical learning theory [15].

3.4 Sequential Mining Optimization

SMO solves the optimization problem when training support vector machines. SMO takes an iterative approach to solve the optimization problem where it breaks problem into the smallest number of sub-problems and solve them analytically.

3.5 Limitation

1. To classify sentiment by creating sentiment lexicon.
2. Unsupervised and semi-supervised learning approaches not give better result than supervised approach for classification.
3. The same word can be used positively as well as negatively, and this difference could only be told by looking at the context.
4. Sentiment classification on traditional topic based categorization.
5. In hybrid approach when PCA integrated with SVM does not provide consistent output.
6. Accuracy problem with naïve Bayes classifiers.

4. Related Work

In 2014 Malhar Anjaria,Ram Mohana Reddy Guddeti proposed influence factor based opinion mining of twitter data. They used supervised learning approaches which combine Support Vector Machine(SVM) and Principle Component Analysis(PCA).This system has achieved good accuracy [17].

In 2012 Balakrishnan Gokulakrishnan, Pavalanathan Priyanthan, Thiruchittampalam Ragavan, Nadarajah Prasath and AShehan Perera presented another approach for sentiment analysis on twitter data stream. They have compared different classifiers such as naïve Bayes, SVM, Maximum Entropy, Sequential Mining Optimization (SMO), Random Forest and Filtered Classifier (FC).They conclude that SMO and Random Forest gave good accuracy. For reducing skewness they used SMOTE technique for data sampling [15].

Kontopoulos, E., Berberidis, C., Dergiades, T., Bassiliades, N. proposed ontology based approach for sentiment analysis of twitter posts. The proposed approach has divided in two phases. First one was creation of domain ontology and second based on the sentiment analysis. They presented Formal Concept Analysis (FCA) and ontology learning approaches [11].

Zhen Niu, Zelong Yin and Xiangyu Kong proposed new model for sentiment analysis based on machine learning and Naïve Bayesian classifier. By using this model they improved the overall efficiency and classified sentiment as positive and negative [16].

In 2011 Hai-Bing Ma,Yi-Bing Geng,Jun-Rui Qiu implemented opinion mining tool which hybrids three different methods such as Semantic Pattern,Semantic Lexicon and KNN or SVM text classification method. They compared these methods and conclued that method based on semantic pattern was good.But if any topic changes frequently then method based on semantic lexicon should be used[18].

Cheng Mingzhi, Xin Yang, Bao Jingbing, Wang Cong and Yang Yixian proposed a graph based sentiment tagging Algorithm for text sentiment classification. Then they used random walk algorithm on the graph and sentiment score has been calculated. They compared SO-PMI and SVM with their methods. They proved that SCG was better than SO-PMI and SVM. So their system improved the efficiency of semantic orientation method but with the use of only context information in sentences [12].

Polpinij, J. & Ghose Presented ontology based method for sentiment classification to classify and analyze the online product reviews. They implemented with Support Vector Machine based on lexical variation ontology. Then sentiment classifier built and tested lexical variations and synonyms in the ontology. They built sentiment classifier based on SVM and achieved more precise result [19].

Pang, L. Lee, and S. Vaithyanathan compared different classifiers such as Naïve Bayes, Maximum Entropy, SVM and various feature extraction methods including unigram, bi-gram and hybrid(unigram+bi-gram).The experimental results show that SVM gave best performance than naïve Bayes [14].

Peter D. Turney proposed unsupervised learning algorithm and reviews have been classified as recommended or not recommended. They employed PMI-IR (Point wise Mutual Information-Information Retrieval) to measure the similarity of words or phrases. They used semantic orientation methods for sentiment classification and achieved good accuracy but problem addressed by using semantic orientation combined with the features in supervised classification algorithm [13].

IV. CONCLUSION

Opinion mining or sentiment analysis has an important role in many areas There are various approaches used for sentiment classification like muchine learning and semantic orientation.Supervised classifiers such as Support Vectore Machine(SVM) , Maximum Entropy,Naïve Bayes,Random Forest and Sequential Mining Optimization(SMO) etc and feature extraction methods like unigram, bi-gram ,hybrid(unigram+bi-gram) are used for sentiment classification,but in most of the cases SVM and hybrid approach for feture extraction can give the good result.Artificial Neural Network(ANN) have been discussed very less number of times.But current opinion mining result has no semantic meaning. We also conclude that if we combine semantic web with sentiment analysis then it can give more precise result.So ontology based approaches could be used to determining subjects discussed in tweets.Our survey suggest that integrating semantic web with web usage mining by using ontology as background knowledge can be useful for sentiment analysis.

V. REFERENCES

[1] B. Pang and L. Lee. "Opinion mining and sentiment analysis". Foundations and Trends in Information Retrieval, 2(1-2):1–135, 1 2008.

[2] Tim Berners-Lee, James Hendler and Ora Lassila .The Semantic Web Scientific American: Feature Article: The Semantic Web: May 2001.

[3] GerdStumme, Andreas Hotho, Bettina Berendt, "Semantic Web Mining State of the Art and Future Directions", ESELIVER, Knowledge and Data Engineering Group, University of Kassel, D-34121 Kassel.

[4] C.S.Bhatia, Dr. Suresh Jain,"Semantic Web Mining: Using Ontology Learning and Grammatical Rule Inference Technique", IEEE, Department of computer engineering, Mewar University, Chittorgarh- 2011

[5] T. R. Gruber."Towards Principles for the Design of Ontologies used for Knowledge Sharing". In N. Guarino and R. Poli,editors, Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer, Netherlands, 1993. Kluwer.

[6] Hakan Yilmaz, "Using Ontology Based Web Usage Mining and Object Clustering For Recommendation", the Graduate School Of Natural And Applied Sciences Of Middle East Technical University, May-2010.

[7] Abd-Elrahman Elsayed1, Samhaa R. El-Beltagy2, Mahmoud Rafea1, Osman Hegazy3,"Applying data mining for ontology building", 1 The Central Laboratory for Agricultural Expert Systems, Giza, Egypt.2 Faculty of Computers and Information, Computer Science Department, Cairo University Giza, Egypt. 3Faculty of Computers and Information, Information System Department, Cairo University Giza, Egypt.

[8] Katarzyna Wójcik, Janusz Tuchowski, "Ontology Based Approach to Sentiment analysis".June-2014.

[9] Sam, K. M. I. Chatwin, C. (2013, Grudzień). "Ontology-Based Sentiment Analysis Model of Customer", International Journal of e-Education, e-Business, e-Management and e-Learning, 3(6), strony 477-482.

[10] Cooley, R. and Mobasher, B. and Srivastava, J. (1997) Web mining: "Information and pattern discovery on the World Wide Web". In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Los Alamitos.

[11] Kontopoulos, E., Berberidis, C., Dergiades, T., Bassiliades, N., "Ontology-based Sentiment Analysis of Twitter Posts", Expert Systems with Applications (2013)

[12] Cheng Mingzhi, Xin Yang, Bao Jingbing, Wang Cong and Yang Yixian. "A Random Walk Method for Sentiment Classification", Second International Conference on Future Information Technology and Management Engineering, 2009.

[13] P. D. Turney. "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews". Proc of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 417-424.

[14] Pang, L. Lee, and S. Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques". Proc of EMNLP-02, the Conference on Empirical Methods in Natural Language processing.

[15] BalakrishnanGokulakrishnan , Pavalanathan Priyanthan , ThiruchittampalamRagavan ,Nadarajah Prasath and AShehan Perera,"Opinion mining and sentiment analysis on a twitter data stream", The International Conference on Advances in ICT for Emerging Regions - ICTer 2012 : 182-188.

[16] Zhen Niu, Zelong Yin and Xiangyu Kong, "Sentiment Classification for Microblog by Machine Learning", 2012 Fourth International Conference on Computational and Information Sciences

[17] Malhar Anjaria and Ram Mahana Reddy Guddeti, "Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning", National Institute of Technology Karnataka, Surathkal, Mangalore - 575025, India.

[18] Hai-Bing Ma, Yi-Bing Geng and Jun-Rui Qiu, "Analysis Of Three Methods For Web-Based Opinion Mining", proceedings of the 2011 international conference on machine learning and cybernetics, guilin, 10-13 july, 2011.

[19] Polpinij, J. & Ghose, A. K. (2008). "An Ontology-Based sentiment Classification Methodology For Online Consumer Reviews", IEEE/WIC/ACM International Conference on Intelligent Agent Technology (pp. 518-524).