Designing an integrated semantic framework for structured opinion summarization

Ehsan Asgarian, Mohsen Kahani

Ferdowsi University of Mashhad, Mashhad, Iran
ehsan.asgarian@stu-mail.um.ac.ir

Abstract. Knowing about people's opinions and viewpoints plays an essential role in decision-making processes involving regular customers to executive managers. Therefore, in the past decade, with the advent of Web 2.0, a new orientation of natural language processing science called opinion mining has been emerged. The main problem of exploring feature-level opinions is the complexity of feature extraction and its relations with the words containing the sentiment within unstructured texts, which reduces the accuracy of opinion mining. The purpose of the structured opinion summarization is to demonstrate the mentioned features in the reviews and express the sentiment value of users for each feature, quantitatively. The main idea of this research is to consider the semantic (knowledge) to analyze the sentiment in the review by developing the opinion ontology. Therefore, a semantic framework as an integrated method is proposed in all stages of feature-based opinion summarization.

Keywords: #eswcphd2014Asgarian · Semantic Framework · Opinion Ontology · Sentiment Analysis · feature based opinion summarization

1 Introduction

One of the newest areas of research on natural language processing, information retrieval and text mining is opinion mining. In general, contextual information can be divided into two sets of facts (explicit information) and opinions (sentiments or implicit information). The primary aim of opinion mining is to extract, classify and summarize people's viewpoints and opinions on various features of an entity or a specific event among valid resources. Most of the work done so far in the field of opinion mining has been on the market and commercial products from the viewpoints of costumers (to select and purchase goods) or distributors (to improve business, competition in the market, effective advertising placement, benchmarking and the recognition of users' tastes and interests). Furthermore, there are applications in medical fields, social science, management and politics. Work on this research area is rapidly growing and new applications of opinion mining in different areas for optimal interactions and decision-making issues of managers or users can be defined.

In general, a sentiment analysis can be classified into three levels including the document level (review), the sentence level (semantic phrases) and feature (aspect)-based level; the feature-based level has been recently taken into consideration by many researchers. Initial studies on opinion mining frequently attempted to classify the opinions or overall sentiments of a document into positive or negative feedbacks [1]. Afterwards, researchers tried to determine the satisfaction or dissatisfaction degree of the document (instead of a two-state classification) [2]. Often supervised methods, in which sample labels are manually marked, were used for these categories in commercial product fields where reviews are directly expressed. The main problem at this level is the assumption that the topic is the same for the entire gathered text or documents. However, different parts of a document (different reviews) may deal with various issues.

Thus, it is vital to identify the topics of different sections and study them separately before analyzing the sentiment. Therefore, opinion mining researchers continuously conducted their work on analyzing the sentiment at sentence level [3] or semantic phrase level [4]. Subjectivity analysis is generally applied to distinguish between subjective sentences and objective ones (e.g. facts such as news) at this level. In recent years most conducted researches in this field have been aimed at non-English languages [5-7]. The major problem of opinion mining at sentence level is due to the assumption that writer's opinion is the same in the entire document. In other words, there can be various opinions (more than one sentiment) on different features (topics) in a sentence. Moreover, in many cases entities (concepts) and their features are not well defined or separated by analyzing the sentiment at sentence level.

Thus, a feature-based approach to opinion mining was proposed owing to existing problems for analyzing the sentiment at document and sentence level [8, 9]. In this approach, entities (topics) and their expressed features are firstly extracted from the text and then the expressed opinions are analyzed for each feature. For example, consider this sentence "Nokia has a good call quality but it is rather expensive!"; Remarks about Nokia cell phone entity (the target) and the call quality and price features are positive and negative respectively.

Compared to simple text summarizers, structured summarization of opinions has been formed according to feature based sentiment analysis, in which useful and relevant information will be available to users. In other words, the purpose of the structured opinion summarization is to demonstrate the mentioned features in the reviews and express the sentiment value of users for each feature quantitatively. The main problem of exploring feature-level opinions is the complexities of feature extraction and their relations with the words containing the sentiment within unstructured texts, which reduces the accuracy of the opinion mining. The following figure shows an example of a summary generator based on opinion features.

In this research, a semantic framework is designed for structured summarization (based on features) of opinions. In the main phase of the proposed framework, we develop the opinion ontology for reviews by receiving opinions within various domains with different languages and, therefore, it can be used for the bulk of reviews. In other words, we can extract features of the text, analyze the sentiment, integrate and summarize opinions by the developed ontology. Using the framework of the pro-

posed ontology, the output results of the structured summarization will be presented as semantic data (e.g. RDF).

2 State of the art

In general, a feature-based opinions summarization comprises three main steps, extracting the features, sentiment analysis and the integration or summarization of them. In the information retrieval area, many methods for extracting concepts and relations between them within the documents have been proposed. However, the purpose of these methods is to identify the main subject of the document and detect the words describing this subject. The relations between subjects are also identified based on their common words and by using various methods of determining the string similarity.

Different methods used to extract features in the review can be divided into five categories: 1) frequent nouns and noun phrases 2) based on relations between the feature and the opinion 3) supervised learning methods 4) topic modeling techniques and 5) hybrid methods. Most initial researches into extracting features from the document were based on nouns and relations between a feature and a sentiment expressions.

The second phase of opinions summarization aims to detect and rank the sentiment regarding each of detected features. Thus, two different approaches are employed: 1) supervised learning approaches 2) Sentiment-lexicon-based approaches.

After extracting the features and determining the sentiment of reviews, obtained results of two previous steps are combined so as to produce a summary of opinions about various features. Hence, similar features in synonymous groups should be merged together and their correspondent sentiments should be averaged. Finally, results in a structured way (quantifying the sentiment for each feature) or selection of positive and negative sentences about each feature in order of preference (time, intensity and …) are displayed.

Semantic approaches have been recently favored by the researchers of the field. This paper focuses on opinion mining methods, exclusively. A new area of semantic techniques for opinion mining with the aim of extracting the features of the main entity in a hierarchical structure and determining the sentiment expressed for each feature has been recently created. However, most of these methods apply the ontology of a particular commercial product which has been manually developed by an expert so as to use the semantic in opinion mining. Generally, none of the previous studies has been worked on converting the reviews into semantic data.

The most significant part of semantic based opinion mining is to create the ontology for a group of opinions. The opinion ontology should extract features, sentiment expressions and the relations between them by receiving the basic and limited knowledge from an expert (the primary list of the sentiment words and the structural taxonomy of features within a desired domain) and applying an automatic method.

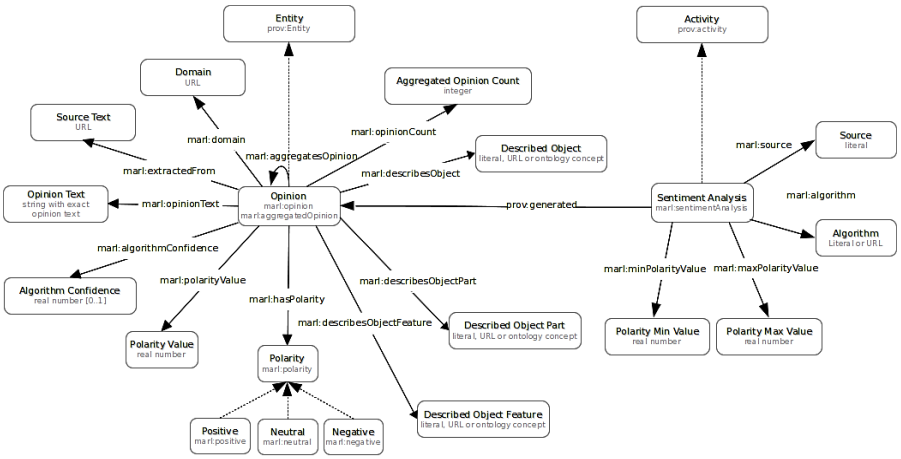

Figure1. Class and Properties Diagram for the Marl Ontology v1.0

2.1 First phase: Ontology schema design

To develop the opinion ontology, the first step is to design a conceptual model or ontology schema of opinions. In fact, ontology schema gives us a logical structure to keep the main entity (main subject) along with its aspects and relevant features.

In [10], a comprehensive model for keeping opinions was proposed as a quintuple (o_j, f_{jk}, so_{ijkl}, h_i, t_l). Where o_j is an object or a main entity (a target object), f_{jk} is a feature of the object, so_{ijkl} is the sentiment value, h_i is the opinion holder and t_l is the time at which the opinion is given. This definition provides a framework to transform unstructured text to structured data. The quintuple above is basically a database schema, based on which the extracted opinions can be put into a database table [11]. Accordingly, semantic data formats have been proposed to keep opinions and the sentiment value of them [12]. Figure 1 shows Marl ontology schema v1.0. However, it is not possible to express comparative opinions or conditional statements in these models. Moreover, features such as date and time (in Marl ontology), trust and data integrity have not been taken into consideration.

In Opinion-ML, a new structure based on XML Schema has been recently proposed according to the developed Emotion-ML model [13]. The main problems of this method are lack of possibility for expressing constraint of parameters, capability to express relationships between opinions and support for comparative opinions.

2.2 The second phase: Development of opinion ontology

Studies have been recently conducted into the usage of domain ontology or product ontology in opinion mining [14, 15]. However in all of them, it is assumed that this ontology is manually given to the system by an expert. In [16], a semi-automated method for developing the ontology of opinions called FDSOT for a specific product has been presented. Nevertheless, in fact, the FDSOT ontology is a bipartite graph

which is simply composed of features and opinions on each one. In order to construct fuzzy domain ontology tree, features are initially identified and the hierarchy of features based on the lexical similarity and the user's knowledge are determined afterwards. However, this ontology depends on the domain, without logical schema and entirely useless for more complicated domains.

In a similar methodology in FCA (Formal Concept Analysis) system [17], some messages are reviewed by the expert and a feature/sentiment expression cross-table is executed manually. Then the ontology of word relations is semi-automatically developed by the OntoGen tool. It has some problems as follows: not using the ontology schema, limited to brief reviews on a specific domain, not considering the relations between sentiment expressions and limited to extracting one type of relation (Sub-Concept-Of).

2.3 Third phase: Converting reviews to the semantic format using the opinion ontology

Various methods for using the pre-built ontology of products to extract features and their sentiment expressions have been proposed [14, 15]. However, most knowledge-based opinion mining methods use the ontology of opinions created by an expert in very few domains. Then they attempt to expand the input ontology and adapt the words within the ontology to the reviews in order to extract features and their sentiment expressions. Moreover, there is no framework of the ontology for mapping the concepts of the opinions and relation between them which is another problem of the current methods. Figure 2 shows an example of the ontology of digital camera for use in opinion mining.

Moreover, a framework for detecting a sentiment was presented using the predefined ontology of products by an expert in [18]. In Kontopoulos's article [17], a method for extracting features of various entities in twitter messages has been presented in terms of ontology, which is used to determine the interest or trends according to features of various products and rank them.

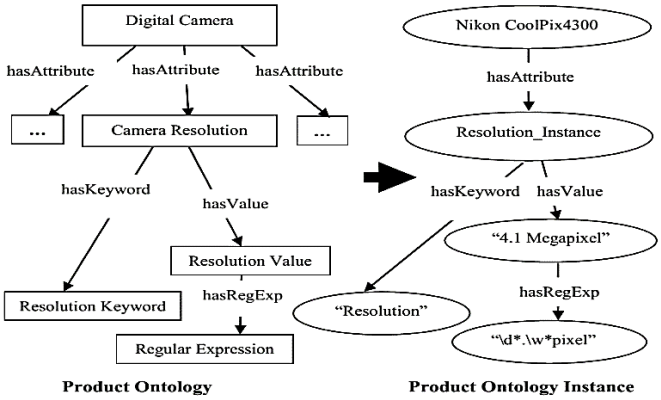

Figure2. Product ontology used in [15] which was created manually.

3 Problem Statement and Contributions

The current feature-based opinion mining methods purely use statistical methods, machine learning technologies or syntactic relations of components for a sentence to automatically extract features and the sentiment expressions. Hence, they have many weaknesses in dealing with linguistic and conceptual complexities to identify the sentiment of opinions. Thus, considering the existing complexities of identifying the entity (the main subject), extracting the features and detecting the sentiment associated with each feature, it is vital to employ semantic methods. Employing a knowledge-based opinion mining method helps to determine various features, the relationship between them and the main entity as well as the expressed sentiment for each feature in complex domains.

Thus, in this research, a semantic framework is proposed to be applied in an integrated method in all stages of opinion summarization. The purpose of this semantic framework is to convert the bulk of opinions into the RDF format (semantic structured information) using the opinion ontology at the reasonable time and applicable to various languages and domains. However, we need to have the full knowledge of various semantic domains so as to develop a general opinion ontology, which is virtually unattainable. Thus, a semi-automatic method is presented to create the opinion ontology in a specific semantic domain. Hence, a conceptual model or ontology schema of opinions is designed to keep opinions in a structured form. Next, given the complexities of natural language to express the sentiment, the target language is determined and the opinion ontology is formed using the opinion documents on a specific domain. Then we can use it to extract features of opinions and detect sentiment expressions for each feature.

Considering the fact that there are a lot of different features, applying this ontology rather than non-semantic approaches results in accuracy improvement and time complexity reduction to extract features and recognize general or feature-specific sentiment expressions in reviews. As a further matter, it is possible to calculate semantic similarities of different features of an entity by the help of various perfectly defined relations in the ontology. As a result, synonymous features are categorized and their correspondent sentiment quantities are combined together. Sentiment quantity is calculated using quantification of sentiment expressions describing features in reviews. Moreover, taxonomic relations defined in the ontology help us determine sentiment quantity used in general and specific features, more accurately.

4 Proposed Approach

As mentioned earlier, the main objective of our research is to propose a semantic framework for using it in the all steps of feature-based opinion summarization. For this end, the opinion ontology is made using a semi-automated method and applying it on domain specific reviews corpus to analyze the sentiment in the new reviews. Therefore, a semantic framework as an integrated method is proposed in all stages of feature-based opinion summarization .

The aim of the proposed framework is to convert the bulk of unstructured reviews into the structured semantic data format in the scalable time as well as being applicable to various languages and domains. Therefore, due to the complexities of the feature-based opinion summarization, in this research, semantic methods are employed to identify the entity (main subject), extract features, detect the sentiment associated with each feature and finally show the relationships and visualize the results. Using a knowledge-based approach helps to determine various features, the relation between them and the main entity as well as the expressed sentiment for each feature in complex domains.

Before developing the opinion ontology, a conceptual model (ontology schema) of opinions independent of the language and the domain is proposed to keep them in the structured format. Next, given the complexities of natural language to express the sentiment, we select the target language and the opinion ontology is formed using the basic knowledge of the user and the domain-specific corpus of reviews. The most significant part of this research is to present an (semi)automatic method for creating the opinion ontology. The quality of obtained ontology plays an important role in accuracy of the proposed structured opinion summarization. The creation of the ontology is completed through three steps in an iterative and incremental process: 1) Extraction of features 2) Sentiment expressions detection 3) Grouping the synonymous features and the determination of the relations between features and the sentiment expressions, with an expert's feedback in an iterative process. By repeating these steps, we can make use of the obtained features and the sentiment expressions from previous iteration, for extracting new ones in the next process. In the meanwhile, the usage of an expert's knowledge (feedback) for the verification of obtained features and the sentiment expressions in each repeat prevents error propagation and improves accuracy. In order to extract the features, an iterative approach is suggested to combine the existing methods and their improved versions.

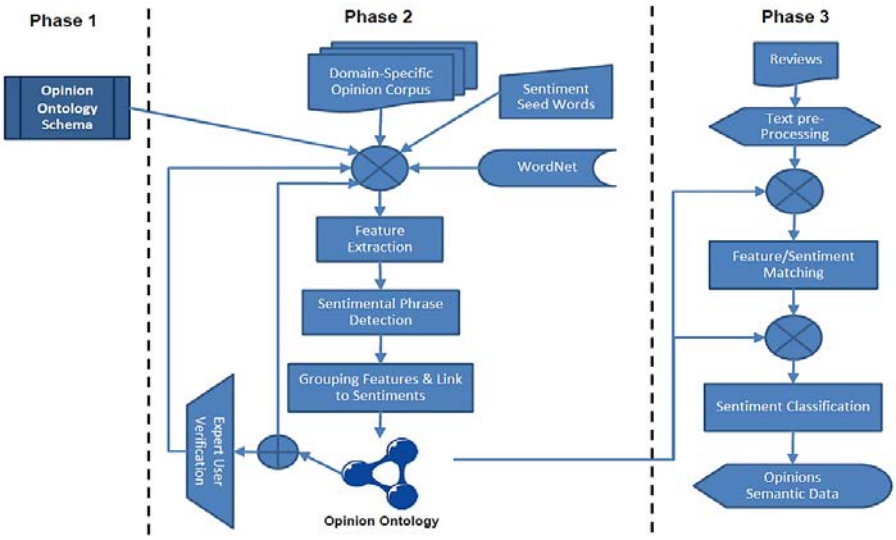

Figure3. Proposed semantic framework for the structured opinion mining

Then the ontology of opinions developed in the previous step will be used to detect the features and various opinions on the desired domain. Using the ontology of opinions, detected features with their sentimental expressions are classified and will be expressed in the form of semantic data (e.g. RDF format). Finally, analysis of the semantic data required to identify the sentiment of the conflicting opinions and essential inferences from comparative ones is conducted. Figure 3 shows proposed semantic framework for the structured opinion mining.

5 Evaluation Strategy

The purpose of this study is to present a semantic framework for the feature-based opinion mining. To assess the accuracy of the proposed method for opinion mining, a labeled test set should be used. However, feature labels and the sentiment values of opinions in an actual data set are not determined and these labels must be prepared by an expert In some commercial data sets the overall rating given to a product by the opinion holder (1 to 5 stars) is used as the sentiment (satisfaction) value of the whole document. In similar methodologies, two methods for measuring the accuracy of the extracted features and estimating error rate of the sentiment of each feature are employed to assess the feature-based opinion mining methods. In order to determine the accuracy of extracted features, a mapping between groups of real features and detected ones based on similarity of words of each feature set (synonyms) is established. Then metrics including the precision, recall and F1-measure are used. Moreover, in order to calculate the accuracy of the predicted sentiment value for each feature, the mean absolute error (MAE) or the mean square error (MSE) measures are used. In some papers [19, 20], the ranking loss measure is used to calculate the accuracy of the sentiment of each feature of various products. This measure demonstrates the average distance between the predicted sentiment and the main sentiment value (assigned by an expert) for each feature which is equivalent to the mean absolute error.
Given that there is no structured semantic opinion summarization system, features and challenges of comparing systems in different stages and phases have been summarized in the following table:

Table 1. Comparing systems in various stages of structured opinion summarization

Phase	Methodology	Domain
(Phase I) Ontology schema design	*Opinion Model*[10]	Domain-independent
	Marl Ontology[12]	Domain-independent
	Opinion-ML[13]	Domain-independent
(Phase II) Creating the opinion ontology	*FCA*[17]	A cell phone on twitter
	FDSOT[16]	Laptop (in Chinese)
(Phase III) Ontology-based approach for opinion mining	*OSPM*[18]	IMDB movies
	Somprasertsri's Method[15]	Cameras
	Mart´Inez's Method[14]	IMDB movies

6 Expected Results

An important outcome of this research is to provide a semantic framework for employing the semantic methods integrated in all stages of opinion summarization. Therefore, using the semantic framework, it is possible to transform the bulk of opinion documents into structured semantic data at reasonable time. Moreover, methods for conducting an analysis of the comparative and conditional opinions, drawing inference about them and combining the conflicting opinions (expressions containing the opposite sentiment) will be presented.

Another advantage of using opinion ontology is to develop relations between concepts and features of opinions and the concepts of other ontology and linked data on the Web. Furthermore, we can use visual tools of the current ontology such as protégé [21], OntoGen [22] and RDF Gravity [23] to express and demonstrate the structured summary of opinions. Thus, in order to convert the opinion documents into semantic data, we can present various categories based on the sentiment (such as the positive and negative points of the entity) or feature (monitoring the opinions on a specific feature) to the user. It is also possible to search and draw better inferences on semantic data in opinions. The ontology schema has to be designed independently of the domain and language, so that it can be used within the various domains of reviews.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86. Association for Computational Linguistics (2002)
2. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 115-124. Association for Computational Linguistics (2005)
3. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. Proceedings of the 2003 conference on Empirical methods in natural language processing, pp. 105-112. Association for Computational Linguistics (2003)
4. Kim, S.-M., Hovy, E.: Extracting opinions, opinion holders, and topics expressed in online news media text. Proceedings of the Workshop on Sentiment and Subjectivity in Text, pp. 1-8. Association for Computational Linguistics (2006)
5. Alm, C.O.: Subjective natural language problems: motivations, applications, characterizations, and implications. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, pp. 107-112. Association for Computational Linguistics (2011)
6. Abdul-Mageed, M., Diab, M., Korayem, M.: Subjectivity and sentiment analysis of modern standard Arabic. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 2, pp. 587-591 (2011)
7. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36-44. Association for Computational Linguistics (2010)
8. Lu, B., Ott, M., Cardie, C., Tsou, B.K.: Multi-aspect sentiment analysis with topic models. Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pp. 81-88. IEEE (2011)

9. Hu, M., Liu, B.: Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177. ACM (2004)
10. Liu, B.: Sentiment analysis and subjectivity. Handbook of natural language processing 2, 568 (2010)
11. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies 5, 1-167 (2012)
12. Westerski, A., Iglesias Fernandez, C.A., Tapia Rico, F.: Linked opinions: Describing sentiments on the structured web of data. 4th International Workshop Social Data on the Web, pp. 10-21. CEUR (2011)
13. Robaldo, L., Di Caro, L.: OpinionMining-ML. Computer Standards & Interfaces 35, 454-469 (2013)
14. Peñalver-Martínez, I., Valencia-García, R., García-Sánchez, F.: Ontology-guided approach to feature-based opinion mining. Natural Language Processing and Information Systems, pp. 193-200. Springer (2012)
15. Somprasertsri, G., Lalitrojwong, P.: Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization. Journal of Universal Computer Science 16, 938-955 (2010)
16. Liu, L., Nie, X., Wang, H.: Toward a fuzzy domain sentiment ontology tree for sentiment analysis. Image and Signal Processing (CISP), 2012 5th International Congress on, pp. 1620-1624. IEEE (2012)
17. Kontopoulos, E., Berberidis, C., Dergiades, T., Bassiliades, N.: Ontology-based sentiment analysis of twitter posts. Expert Systems with Applications 40, 4065-4074 (2013)
18. Zhou, L., Chaovalit, P.: Ontology supported polarity mining. Journal of the American Society for Information Science and technology 59, 98-110 (2008)
19. Lu, Y., Zhai, C., Sundaresan, N.: Rated aspect summarization of short comments. Proceedings of the 18th international conference on World wide web, pp. 131-140. ACM (2009)
20. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. Proceedings of the 17th international conference on World Wide Web, pp. 111-120. ACM (2008)
21. Noy, N.F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R.W., Musen, M.A.: Creating semantic web contents with protege-2000. Intelligent Systems, IEEE 16, 60-71 (2001)
22. Fortuna, B., Grobelnik, M., Mladenic, D.: OntoGen: semi-automatic ontology editor. Springer (2007)
23. Goyal, S., Westenthaler, R.: Rdf gravity (rdf graph visualization tool). Salzburg Research, Austria (2004)