Ontology Supported Policy Modeling in Opinion Mining Process*

Mus'ab Husaini[1], Andrea Ko[3], Dilek Tapucu[1,2], and Yücel Saygın[1]

[1] Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul, Turkey
[2] Dept. of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey
[3] CUB, Hungary
{musabhusaini,dilektapucu,ysaygin}@sabanciuniv.edu,
andrea.ko@uni-corvinus.hu

Abstract. In e-Society the spreading services offered by Social Web has changed the way of communication and cooperation among citizens, policy-makers, governance bodies and civil society actors. One of the main goals of policymakers is to motivate citizens for participation in policy-making processes. UbiPOL ((Ubiquitous Participation Platform for Policy-making, ICT-2009.7.3(ICT for Governance and Policy Modelling), 2009-2011) aimed to develop a ubiquitous solution, which emphasizes citizens' participation in policy-making processes (PMPs) regardless of their current location and time. Ontology-based opinion mining component of Ubipol system has a crucial role in citizens' commitment, because it empowers them to contribute in policy making. This paper presents the ontology-based semi-automatic approach and tool for sentiment analysis in Ubipol system, which include lexicon extraction from a large corpus of documents. Aspect-based opinion summarization of user reviews and its combination with domain ontology development are discussed as well.

1 Introduction

Ubiquitous Participation Platform for Policy-making (UbiPOL) project aimed to provide a ubiquitous participation platform that allows citizens to participate in the policy making process during their everyday life through providing relevant policies and others opinions that a affect their life wherever they are located. The specific objectives of UbiPOL are to [6]:

- develop an executable policy making process model that is related with geography through,
- geocode policy issues,
- attached policy issues to existing site objects (point of interests, POIs) and
- track the policy making process which is formed by citizen's opinions according to citizen's location and input opinions

* This work was developed in the context of UBIPOL (Ubiquitous Participation Platform for Policy Making) project funded by European Commission, FP7.

P. Herrero et al. (Eds.): OTM 2012 Workshops, LNCS 7567, pp. 252–261, 2012.

It provides context-aware knowledge provision with regard to policy-making. Citizens using UbiPOL will be able to identify any relevant policies and other citizen's opinions whenever they want and wherever they are, in accordance with and fitting in with their as-usual life pattern. With the platform, citizens are expected to become more widely aware of any relevant policies and PMPs for involvement during their as-usual life; thus, there will be improved citizens' engagement and empowerment. Also, the platform provides policy-tracking functionality via a workflow engine and opinion tag concept to improve the transparency of policy-making processes. Ubipol system enables policy-makers to collect citizen opinions more efficiently as the opinions are gathered as soon as they are created in the middle of the citizens' usual life. UbiPOL provides security and an identity management facility to ensure that only authorised citizens have access to the relevant policies. UbiPOL services are provided through a scalable platform ensuring that a large number of citizens can make use of the system at the same time (for example, for e- Voting applications) via its well-proven automatic load balancing mechanisms. The privacy-ensuring opinion mining engine prevents the unwanted revealing of citizen identities and the mining engine stops any unrelated commercial advertisements from being included in the opinion base, to minimise misuse of the system.

Ontologies have a crucial role in Ubipol system, they help to structure the policy related context, provide conceptualization for policy domain and used in the opinion mining process. In order to create an ontology that is useful for sentiment analysis, we have to cover a large set of opinion documents from a certain domain. In this process, a person (domain expert) defines all the domain concepts (aspects) and corresponding features from the corpus, which is extended and updated continuously through the opinion processing. Aspect is a key term in our opinion mining solution; it describes certain characteristics of a domain, like environmental issue in policy issues domain. In this paper, we present an ontology based opinion mining engine to enrich policy modeling process (Figure 1). This engine analyzes a domain- specific opinion corpus, meanwhile it is assisting the user with the updating of a domain ontology and then determines the polarity of opinion on the various domain aspects.

In this engine, the context of a word a effects its meaning, specially whether it has a positive or negative or even neutral orientation. Before processing the text to determine its sentiment orientation, the policy domain aspect has to be identified (namely which policy category is represented by the concept). This identification is supported by the policy modelling ontology in Ubipol system, which describes the most important policy - related classes (aspects) and structure. To identify those policy issues, which require special attention is a key goal of policy maker. Ontology-based opinion mining highlights the most important issues, their aspects; polarity (positive, neutral or negative attitude of writers) and life cycle for the decision makers and citizens, through the continuous analyses of opinions and comments. During this process, several research questions emerge. How the set of aspects domain specific features of opinions (aspects) can be obtained? How they are linguistically expressed? How they are related

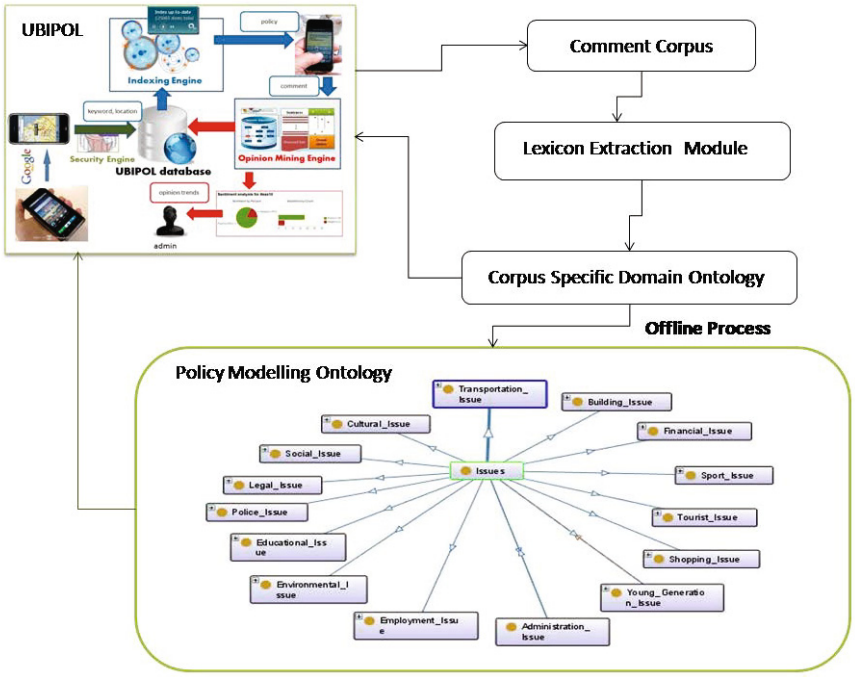

Fig. 1. UBIPOL Architecture and Our Approachthe Ontology-based Opinion Mining Solution

to each other? According to the questions above, the main research questions discussed in our paper are the following: How can we identify and monitor those policy issues, which require special attention or immediate action using citizens' opinions? What kind of algorithms can be applied to extract aspects (decisive characteristics of domain specific opinions) in a reusable manner? How can we structure these characteristics to gain semantic layer (ontologies) which can be integrated with the policy making process?

To answer these questions, we propose an ontology based opinion mining solution. Our main contribution is the ontology based opinion mining process and tool, more precisely the combination of Eagerly Greedy Set Cover algorithm (discussed in detail in section 4) with domain ontology development and maintenance. We aim to identify the smallest set of documents in a given corpus which provide appropriate information needed to develop domain ontology.

2 Related Work

This section gives a literature overview about the main areas; feature extraction and review summarization related to our research challenges. Using feature extraction for sentiment analysis has been studied extensively. In this area, Hu [5] introduced a technique that uses association rules and the most frequent nouns in a given set of reviews. Based on these rules, a set of features can be synthesized

for the domain. In another work [8] a similar concept is used to find frequent noun phrases from the review dataset and then extract the product's parts and properties based on point-wise mutual information scores between these phrases and meronymy descriptors related to the product are found. These technique only focus on the overall features and not the keywords associated with them. In wider domains, this approach will yield large feature sets which are not necessarily useful for aggregation and summarization. In this work, we assume that the real set of features is limited and that various keywords are used to represent these features in opinion documents. The method proposed in [4] takes into account the relationship between a term and its related opinion information. Blair-Goldensohn et al. [1] have also reported a sentiment summarizer with aspect information for local service reviews. In [3] and [7], the authors suggested using a clustering algorithm for aspect identification.

A supervised method which combines frequency, syntax tokens, and domain knowledge to find the product features has been used in [12] to extract product features. The induction of domain knowledge aims to improve the quality of extraction. Gupta and Lehal [2] propose a keyword extraction method to support topic detection and document summarization.

3 Problem Definition and Theoretical Background of Our Research

The main objective of our research is to discover the smallest set of documents in a given corpus to provide appropriate information needed to develop a domain ontology. Before proceed with an explanation of our approach, it will be helpful to outline assumptions, define the problem, and disambiguate the terms commonly used in this paper.

To simplify our problem, we will assume that the corpora are domain-specific, i.e., all opinion documents in a given corpus are concerned with only one domain. An ontology, in turn, contains relationships among the domain, its *aspects*, and their respective *keywords*.

Definition 1 (Ontology). An ontology is a triple of the form, O = (D;A;K), where D is the corpus domain, A is the collection of aspects in domain D, and K is the collection of all keywords in domain D.

Definition 2 (Aspect). An aspect instance is the tuple, A = (a;KA), where a is a noun that denotes a certain characteristic of the domain that can be subject to opinion and KA is the collection of keywords used to represent aspect A. Aspects are also termed to domain features.

Definition 3 (Keyword). A keyword k , is a noun that can be used to represents a given aspect, A, in an opinion document.

An opinion document can then regarded as a set of words containing both keywords and non-keywords. Definitions mentioned above, we can also formalize the definitions of corpora and opinion documents.

Definition 4 (Corpus). A corpus is a tuple of the form, C = (D;R), where D is the corpus domain and R is the collection of opinion documents in the corpus.

Definition 5 (Opinion Document). An opinion document (or simply document) instance, R, is a set of m words such that,

$$R = \{w_1, w_2, w_3, ..., w_m \mid m > 0\}$$

Having defined the problem, we now turn to outlining our approach which consists of approximating keywords and using a clustering technique to find the subset of the corpus that minimizes the number of documents required and maximizes the amount of information available to create an ontology.

Our contribution is to provide a user-friendly and minimal-effort environment for producing gold-standard domain aspect lexica and better understanding of the issues and problems. Our approach is primarily intended for use by sentiment analysis researchers to study aspect lexicon generation and sentiment analysis techniques, and to a certain extent, practitioners. It can also be used by domain experts to study aspect lexicon generation.

4 Ubipol Approach for Opinion Mining

In sentiment analysis, the context of a word affects its meaning, specifically whether it has a positive or negative or even neutral orientation. Before processing for orientation, the concept that the word represents has to be known, which can be defined with a concept ontology. In order to create a robust ontology that is useful for sentiment analysis, we have to cover a large set of opinion documents. In this process, a person (domain expert) possessing domain knowledge defines all the domain concepts (aspects) and corresponding features from the corpus, which becomes increasingly expensive as the size of the corpus increases.

Our tool is designed as a self-contained web application that can be deployed to any Java-based web server. This system provides an interactive interface for users to import a domain corpus into the application in a variety of formats including text and XML. This corpus is then analyzed to obtain the most informative corpus documents from which the user extracts domain aspects and related keywords to create the aspect lexicon. The user also has the option to upload a pre-generated lexicon instead of manually extracting it. The lexicon, manually extracted or user-provided, is then used by the sentiment analysis engine to process the corpus. Finally, results are displayed to the user for further analysis. Fig.2 and Fig.3 illustrates this workflow with the help of screenshots. This system is publicly accessible online by visiting http://ferrari.sabanciuniv.edu/sare, and a demo video explaining how to use the system can be downloaded from the same address. The major components of the system are described below.

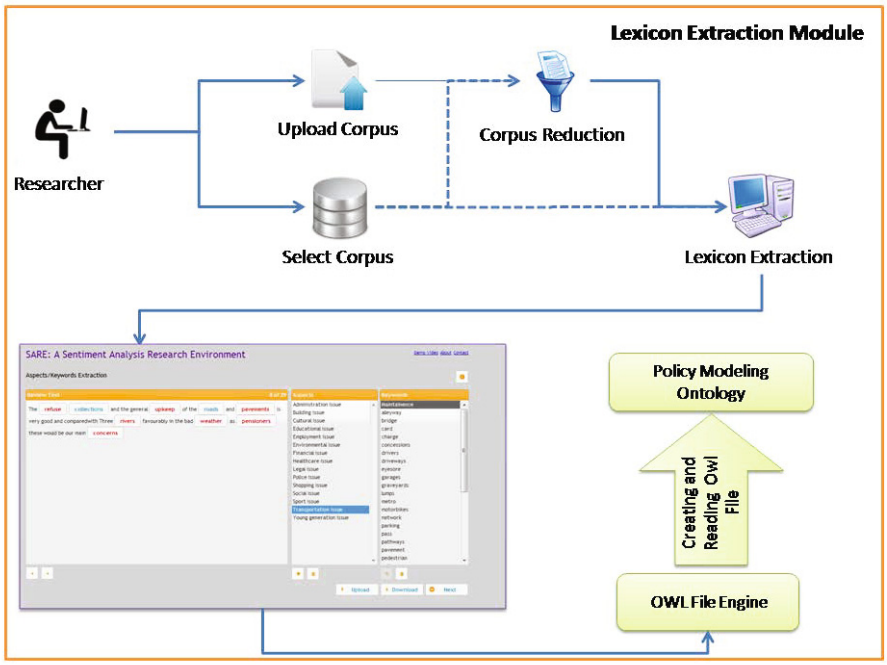

Fig. 2. Lexicon Extraction Module

4.1 Corpus Reduction

This module deals with the problem of aspect lexicon extraction by corpus summarization and user annotation. We approximate aspect keywords with corpus nouns and apply Eagerly Greedy Set Cover algorithm (a variation of the *Greedy Set Cover* algorithm, that we developed) called *Eagerly Greedy Set Cover* algorithm to find the minimum set of documents that cover all nouns in the corpus.

Eagerly Greedy, is given as follows: We maintain a candidate set cover initialized to an empty set, and iterate through the document collection R sequentially. For each document encountered, we consider the set of all nouns in the document and attempt to sequentially consume its elements into the candidate cover sets, i.e., members of the candidate set cover. Each time a candidate cover set consumes (presumes covered) an element, it increments its own utility score by one and removes the element from the new set. If the candidate cover set is a subset of the new set, then the candidate cover set is itself entirely consumed and replaced by the superset (with utility score being set to the sum of existing utility plus the size of the new set). This process continues until either all of the elements in the new set are consumed or we run out of candidate cover sets to process. In the latter case, a new candidate cover set is formed from the contents of the uncovered nouns with the utility score being initialized to its size. It should be noted that while the composition of sets may be changed changed

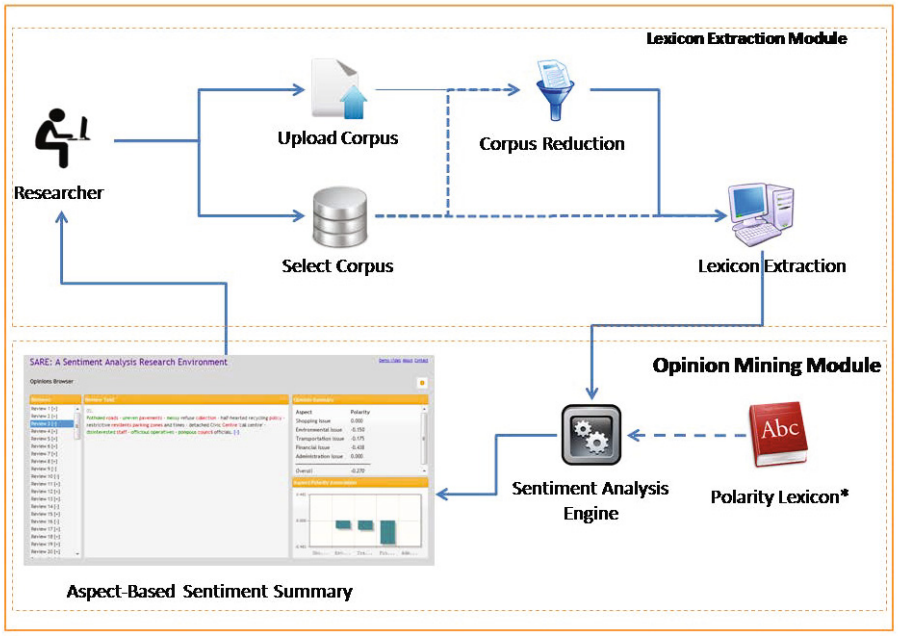

Fig. 3. Opinion Mining Module

by the end of each iteration, each candidate cover set maintains its identity such that the original set can be recovered as needed.

In order to evaluate the extent of data reduction obtained by the *Eagerly Greedy Algorithm*, we need to compare it with the best and the worst approximations of set cover problems. In this case, the worst case approximation is the random selection approach; that is to say, randomly choosing sets from the collection until all the elements in the universe are covered. The best approximation, as referenced previously, is the classical greedy algorithm.

Since our overall goal is to reduce the amount of data presented to the domain expert, we observe the performance of these three algorithms as a function of the amount of data reduction achieved.

Results. The amount of data reduction achieved by each of the algorithms is given in Table 1. We also experimented with several values of $\hat{\tau}$ to observe the value that provides us with the best reduction while incurring the least amount of loss in utility. As shown by the plot of $\hat{\tau}$ against $\Delta(EG_{\hat{\tau}})$. Thus, we can achieve very high utility coverage with a smaller part of the corpus by allowing for some outlier documents to be ignored. The comparison of algorithms shows that while the reduction achieved by the greedy algorithm is slightly higher than our algorithm (a difference of 2.81%), we can leverage the pruning capabilities of our algorithm to dramatically widen the gap in the opposite direction by

Table 1. Algorithm Comparison

Algorithm [α]	Data Reduction [$\Delta(\alpha)$]
Random	0%
Greedy	68.2%
Eagerly Greedy ($\hat{\tau} = 0\%$)	65.39%
Eagerly Greedy ($\hat{\tau} = 8\%$)	96.72%

introducing a small tolerance to error – the error tolerance of 8% that we have chosen for this corpus gives our algorithm an edge of 28.52% over the greedy algorithm.

4.2 Lexicon Extraction

We were also interested in employing a domain expert to discover features from the corpus. Since we do not have a standard ontology for our corpus domain, we had to evaluate the result qualitatively rather than empirically.

Setup. We developed a user interface to facilitate identification of features from provided examples. In this application, the user is sequentially provided with examples from the reduced set for annotation. To help the user easily spot keywords from the larger text, we also highlight all the nouns in a given document. An illustration of a hotel review as displayed in the application is given in Fig. 2. Based on the review that the user is shown, they can then add aspects and related keywords to their respective lists as shown in the same figure.

Our opinion corpora were drawn from a set of *TripAdvisor*[1] hotel reviews as published in [11]. This dataset consists of 235,793 reviews on various hotels that were aggregated over a one month period. For each experiment, we sampled a consistent but random subset of these reviews as will be detailed under each experiment. The experiments were performed using a Java implementation, and document nouns were extracted using the Stanford POS tagger presented in [9].

Results. The application detailed above was used to create an ontology for the any domain from the generated in the earlier steps. From this experiment, we determined that the domain has several aspects, and also obtained several keywords in each of these aspects. A breakdown of domain aspects and number of corresponding features has been provided in Table 2.

4.3 Opinion Mining Module

The objective of this module is to use ontology based approach for representing corpus-specific knowledge and to present aspect-based score value to summarize results in a scorecard structure. Our contribution is to provide a domain-independent approach in a user-friendly fashion. Thus, firstly corpus-based

[1] Online hotel booking and reviews site [10].

Table 2. Keyword Breakdown By Aspect for the Hotel Domain

Aspect	Number of Keywords
Business service	34
Check in/front desk	31
Cleanliness	42
Location	105
Rooms	138
Service	147
Value	23
Total	520

aspects and aspect-related keyword sets are extracted as described in Sec. 4.2 to create a corpus- specific ontology. Then a polarity ontology is created from SentiWordNet [www.sentiwordnet.com]. Finally, polarity-placement algorithm is used to calculate score values for each aspect. The idea of the algorithm is to get initial polarity value from the polar- ity ontology for any opinion word in a given comment and to transfer polarity value from polarity keywords to aspect- related keywords by using the Stanford NLP API. The reason for using the Stanford NLP API is to generate dependency tree graphs for a given sentence. After polarities are transferred on a correct token, aspect-based score value is calculated for each comment. A detailed aspect-based sentiment summary is displayed as illustrated in Fig. 3.

5 Conclusion and Future Work

This paper presented an ontology-based semi-automatic approach and tool for sentiment analysis in Ubipol system. The problem of lexicon extraction from a large corpus of documents was discussed with a focus on the aspect-based opinion summarization of user reviews. The main components of our solution are the lexicon extraction module, feature extraction module, which are detailed in section four.

The main challenge was to identify the smallest set of documents for a given corpus that provide appropriate information for domain ontology development. Eagerly Greedy Set Cover algorithm was suggested as a solution, because of its advantageous characteristics. In section four we showed that it provides better results in data reduction than the other approaches, like random and Greedy algorithms. Another advantage of our opinion mining process is that it supports domain ontology maintenance through the continuous processing of citizens' opinions. Policy modelling ontology can be utilized as an aspect/keyword set and fine-tuned in the steps of opinion mining. In spite of the fact, that we offer a user-friendly environment the main target group of the solution include sentiment analyses researchers, practitioners and domain experts. In wider context opinion mining component of Ubipol system supports policy makers to monitor citizens' opinions and they get immediate feedback if special attention or action is needed in a certain policy area.

Future work involves comparison the accuracy and the computational efficiency of our aspect extraction method with other approaches. In addition, larger scale quantitative evaluation of our opinion mining method will be conducted.

References

1. Blair-goldensohn, S., Neylon, T., Hannan, K., Reis, G.A., Mcdonald, R., Reynar, J.: Building a sentiment summarizer for local service reviews. In: NLP in the Information Explosion Era (2008)
2. Gupta, V., Lehal, G.: A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence 1(1) (2009)
3. Hadano, M., Shimada, K., Endo, T.: Aspect identification of sentiment sentences using a clustering algorithm. Procedia - Social and Behavioral Sciences 27(0), 22–31 (2011), http://www.sciencedirect.com/science/article/pii/S1877042811024062; Computational Linguistics and Related Fields
4. Hana, J., Dongwook, S., Joongmin, C.: Ferom: Feature extraction and refinement for opinion mining. ETRI Journal 33(5), 720–730 (2011)
5. Hu, M., 0001, B.L.: Mining opinion features in customer reviews. In: McGuinness, D.L., Ferguson, G. (eds.) AAAI, pp. 755–760. AAAI Press/The MIT Press (2004), http://dblp.uni-trier.de/db/conf/aaai/aaai2004.html#HuL04
6. Irani, Z., Lee, H., Weerakkody, V., Kamal, M.M., Topham, S.: Ubiquitous participation platform for policy makings (ubipol): A research note. IJEGR 6(1), 78–106 (2010)
7. Ly, D.K., Sugiyama, K., Lin, Z., Kan, M.Y.: Product review summarization based on facet identification and sentence clustering. CoRR abs/1110.1428 (2011), http://dblp.uni-trier.de/db/journals/corr/corr1110.html#abs-1110-1428
8. Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 339–346. Association for Computational Linguistics (2005)
9. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL 2003, pp. 173–180 (June 1, 2003)
10. The TripAdvisor website (TripAdvisor LLC) (2011), http://www.tripadvisor.com
11. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis on review text data: A rating regression approach. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 783–792 (2010), http://portal.acm.org/citation.cfm?id=1835903
12. Zhang, S., Jia, W., Xia, Y., Meng, Y., Yu, H.: Product features extraction and categorization in chinese reviews. In: ICCGI 2011, The Sixth International Multi-Conference on Computing in the Global Information Technology, pp. 38–42 (2011)