Towards Opinion Mining Through Tracing Discussions on the Web

Selver Softic and Michael Hausenblas

Institute of Information Systems & Information Management,
JOANNEUM RESEARCH, Steyrergasse 17, 8010 Graz, Austria
firstname.lastname@joanneum.at

Abstract. This paper reports on our ongoing work regarding opinion mining from Web-based discussion forums in the realm of the Understanding Advertising (UAd) project. Our approach to opinion mining is to first RDFise discussion forums in SIOC, and in a second phase to interlink the so created data with linked datasets such as DBpedia. We are confident that this should allow a market researcher to formulate queries using domain semantics and hence understand what people think about a certain product or service. The system's architecture, preliminary results, and the current available demonstrator are discussed in this work.

1 Introduction

Products or services are often discussed by customers on the Web. Whilst (official) company sites usually tell a certain side of the story, having users discussing advantages or issues with certain products offers a source for a deeper understanding of a market. Equally found in common social life, the communities on the Web can have a strong impact on trend setting. This observation is eligible, if we behold a trend as a spot where subjective views of different people lead by their affections may cross and merge with each others.

We found that valuable data relevant for market research on the Web is neither easy accessible nor processable. Time expenses to collect and evaluate data needed for a better market understanding are still tremendous. As recently pointed out by Peter Mika (Yahoo! Research) [1]:

> Current search technology is unable to satisfy any complex queries requiring information integration such as analysis, prediction, scheduling, etc. An example of such integration-based tasks is opinion mining regarding products or services. While there have been some successes in opinion mining with pure sentiment analysis, it is often the case that users like to know what specific aspects of a product or service are being described in positive or negative terms and to have the search results appear aggregated and organized.

In the Understanding Advertising (UAd) project[1] we aim at developing a methodology allowing a market researcher to understand a certain market.

[1] http://www.sembase.at/index.php/UAd

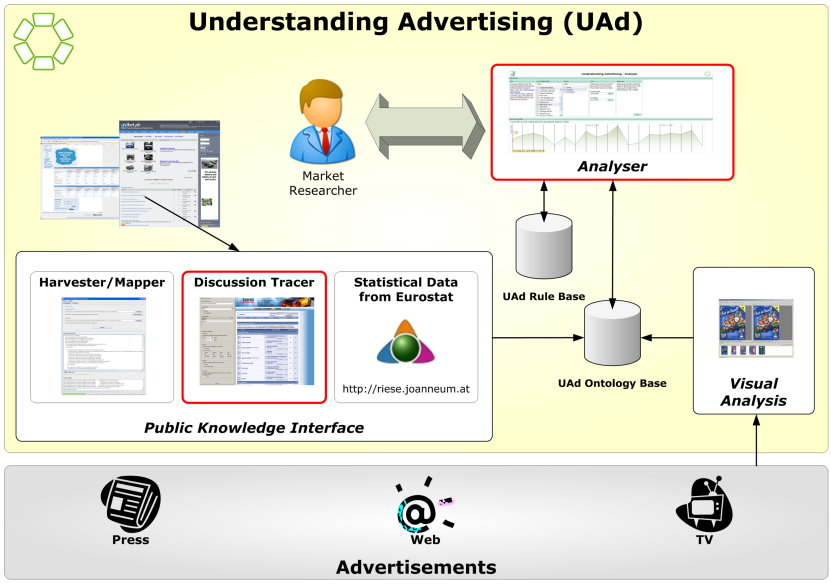

Fig. 1. The UAd system architecture.

The analysis performed in UAd is twofold, (i) by visual interpretation of advertisements (from print media, Web and TV), and (ii) by using information available on the Web. Fig. 1 depicts the overall UAd system architecture, consisting of (i) the UAd Analyser (the front-end for the end-user), (ii) the "Public Knowledge Interface", and (iii) the Visual Analysis module. Information about products and services are gathered from the Web through the so called UAd "Public Knowledge Interface" (PKI). We have developed three methods converting plain (HTML) Web content into structured data represented in RDF allowing us to be both flexible and comprehensive:

1. Plain old screen scraping (in the so called UAd Harvester/Mapper module);
2. Pattern-based RDFising and Interlinking for online discussions (the UAd Discussion Tracer);
3. Schema-based a-priori RDFising and Interlinking (for statistical data from Eurostat; described elsewhere [2, 3]);

In this paper we focus on tracing discussions on the Web, hence the two components involved in this task (Discussion Tracer and Analyser) are highlighted in Fig. 1.

This paper is structured as follows: First, we review related work in section 2. Then, in section 3 we discuss our approach representing discussions and opinions. In section 4 we present the system's architecture, discuss the data acquisition and the market researcher's interface. We present preliminary results in section 5. Finally, we discuss our findings and highlight future work in section 6.

2 Related Work

Recent research on opinion mining has focused on sentiment analysis, simple "pro" and "cons" classification [4] and determination of semantic orientation in opinion models using feature-based opinion summarisation on word, sentence or document level. Typically, Natural Language Processing (NLP) [5–7] and machine learning techniques [4, 8] have been utilised in supervised or unsupervised modes [9, 10] allowing the extraction and classification of sentiment and opinions polarisation. The workflow usually comprises three major phases: extraction, structuring and summarisation of results. In general we subscribe to this pattern, however differ in a number of details mostly regarding the explicit representation of the information.

Motivated by earlier experiences [11, 12] our approach is based on Semantic Web technologies (RDF, SPARQL, etc.). Further, in contrast to existing work, we use widely deployed vocabularies—e.g. Semantically-Interlinked Online Communities (SIOC)—along with existing APIs [13] for the extraction and structuring phase. Regarding the formal representation of products and their characteristics it is worth noting that the W3C has recently launched the "Product Modelling Incubator Group"[2] aiming at creating a product modelling ontology.

We aim at orienting the opinion holder context on domain semantics [8] along with exploiting linked datasets (such as DBpedia [14]) and domain delimited query expansion [15]. Furthermore, the creation of opinion ranking primary for sentiment classification [16, 17] will be considered in greater detail in our future work.

To the best of our knowledge there exists no other work in the area of opinion mining that deals with explicitly modelled opinions along with linked data sets for its domain knowledge.

The basic idea of linked data was outlined by Sir Tim Berners-Lee [18] in 2006. The Linking Open Data (LOD) community project[3] is an open, collaborative effort applying the linked data principles. It aims at bootstrapping the Web of Data by publishing datasets in RDF on the Web and creating large numbers of links between these datasets [19]. The datasets included in the project are diverse in both nature and size. Currently, the project includes some 30 different datasets, ranging from rather centralized ones (such as DBpedia [14]) to those that are very distributed (for example the FOAF-o-sphere). While some of the datasets focus on certain domains (for example the Eurostat data [3]), others are more of a generic type, such as Revyu.com [20].

3 Representing Discussions and Opinions

To support a market researcher in analysing a certain market, one of the sources used in the UAd PKI are Web-based discussion forums. For enabling structured

[2] `http://www.w3.org/2005/Incubator/w3pm/`
[3] `http://linkeddata.org/`

queries and browsing it is necessary to represent the discussions in a machine-interpretable way and enhance it with domain semantics. Web-based discussion forums offer a well-structured source for this purpose, hence the idea to exploit them along with linked datasets.

Our goal is it to explicitly model the opinions in a discussion being compliant to the Web of Data. We decided to reuse an existing vocabulary to represent the discussions rather than reinventing the wheel. Due to its popularity and wide-spread use, the Semantically-Interlinked Online Communities (SIOC) vocabulary[4] has been selected to represent discussion threads and posts.

However, in case of explicitly representing opinions we did not manage to find an appropriate vocabulary. Although one could for example use a review vocabulary[5] as a base and extend it, we found it better suited to define a dedicated vocabulary for this task.

Our "Opinion Mining Core Ontology"[6] (cf. Fig. 2) basically defines the following classes and properties:

- `opm:DiscussionOpinion`, the central hub that connects discussion threads with opinions about a certain entity;
- `opm:Opinion`, an abstract representation of an opinion;
- `opm:Topic`, a proxy concept to trigger aspects of a certain topic.

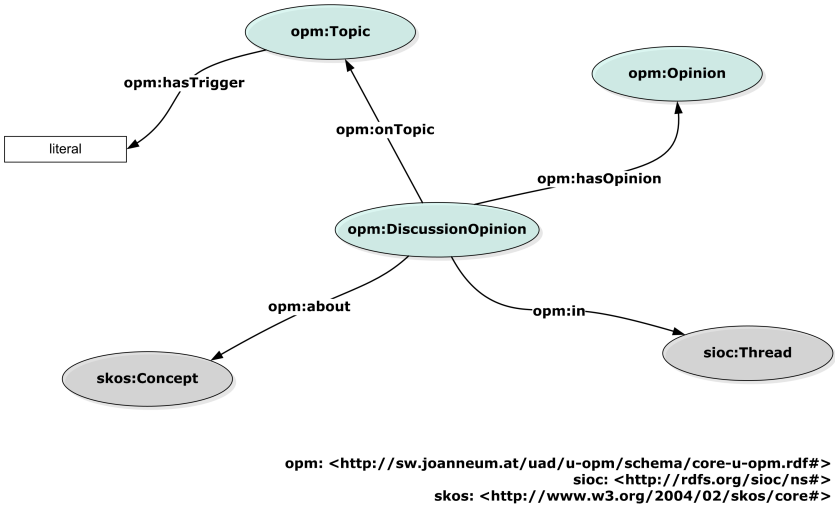

Fig. 2. UAd's Opinion Mining Core Ontology.

[4] `http://www.sioc-project.org/ontology`
[5] Such as `http://danja.talis.com/xmlns/rev_2007-11-09/index.html`
[6] `http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf`

We use `skos:Concept` of SKOS [21] to represent what a discussion *is about*, for example, a certain car such as the Alfa Romeo 156; we note that this design decision also supports the straight-forward utilisation of data from DBpedia. Further, we use the `sioc:Thread` from the SIOC vocabulary to indicate *where* the discussion has been taken place.

It has to be noted that `opm:Opinion` is currently deliberately underspecified. We intend to extend and refine this part to the ontology based on our experiences with the system and regarding earlier work from [9, 10]. Further, we want to point out that the `opm:Topic` concept is used to represent a certain aspect of a discussion, that is, it might indicate that users discuss about the pricing, about problems with a certain product or simply express their satisfaction. The semantics of this concept are such that if one of the assigned trigger words has been found in a discussion, the topic is believed to match (hence the labelling of the datatype property `opm:hasTrigger`).

The introduced lightweight ontology above plays a decisive role in our opinion mining process. In order to achieve better scalability and reusability, it acts as a nexus between the domain of concern and the RDFised data. This is why it makes no difference for our opinion mining model if there is the DBpedia categorisation behind or some other domain specific ontology. Therefore, our approach offers flexibility by choice of domain and yields a generic opinion creation.

4 Discussion Tracing

In the process of discussion tracing in UAd, two major components are involved (Fig. 3), namely (i) the UAd Data Acquisition (highlighted), where Web-based discussions are harvested, RDFised and interlinked, and (ii) the UAd Analyser, allowing to query and access the data.

4.1 Data Acquisition

The data acquisition in UAd is performed in three phases; in a first step the most common data in a Web-based discussion forum, such as title, author, creation date, etc. is RDFised using SIOC. In a second phase the entities occurring in the discussion posts are identified and interpreted regarding a certain domain (in our demonstrator this domain is "cars"). This second step involves the interlinking to linked datasets such as DBpedia or instances of some other domain specific ontology. For our purposes DBpedia offers enough adequate instances and well formed domain model respectively area of interest. However, as mentioned in Section 3 this is not mandatory and DBpedia can be easily replaced by any other domain ontology. Currently, interlinking with DBpedia is done manually, however in the final version we are aiming to automate this task. In a third phase the (subjective) statements of participants are analysed, and further added to the knowledge base. This is mainly achieved by the creation of `opm:DiscussionOpinion` instances and their respective properties. To this

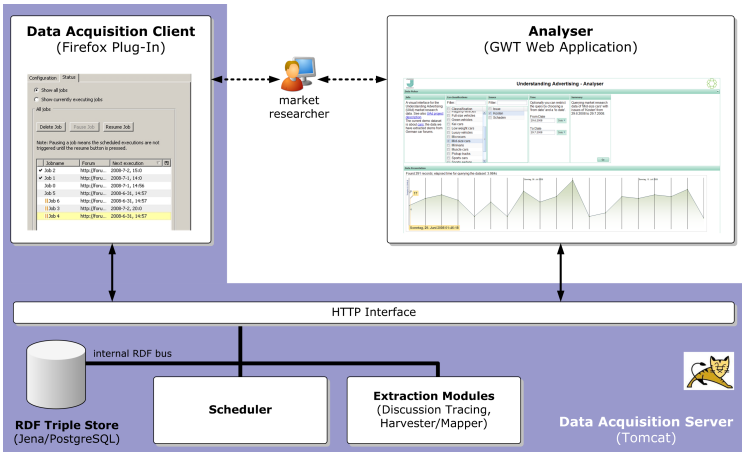

Fig. 3. Discussion Tracing in UAd.

end, we use a manually pre-configured list of possible topics, that is instances of `opm:Topic` to trigger the creation of opinions.

We have implemented a client/server system (Fig. 3, left and bottom) to perform the data acquisition in UAd. Within the scope of our research we support RDFising popular discussion forum types[7] such as vBulletin or phpBB. Data extraction occurs automatically using extracion profiles, manually defined for several forum types; a single acquisition task represents a single job on the server. The server has been implemented using a Java application server (Tomcat) along with a Jena 2/PostgreSQL RDF store taking care of the scheduling and execution of the acquisition tasks.

At the client side, a Firefox plug-in (Fig.4) allows a user to define, control and monitor the tracing tasks. The plug-in has been developed in JavaScript and XUL[8]. A user typically adds the link of a discussion forum and selects the forum type. Currently, only entire forums can be extracted. We plan to support the selection of sub-forums independently from each other for the extraction task. The user can also specify time parameters for the acquisition tasks, for example how often per week a job should be triggered to update the store.

4.2 Analyser

The UAd Analyser is a Web Application allowing a market researcher to examine the data gathered by the UAd Acquistion Server. In Fig. 5 the current state of the implementation (implemented with the Google Web Toolkit[9]) is depicted. The user can limit the data by selecting certain car classifications and issues

[7] `http://www.big-boards.com/statistics`
[8] `http://developer.mozilla.org/en/docs/XUL`
[9] `http://code.google.com/webtoolkit/`

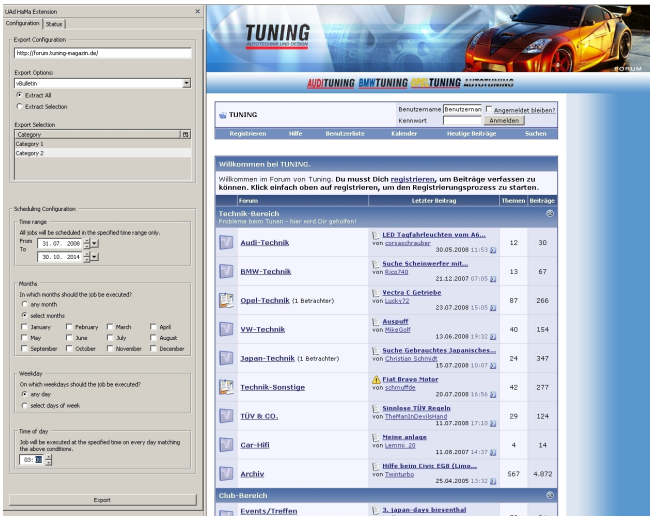

Fig. 4. The DT Plugin.

as well as by restricting the time period. The queried data is visualised with a Simile Timeplot[10] module, displaying the time on the X-axis and the number of discussion posts in the Y-axis. Discussion threads are illustrated as red vertical lines; the users may retrieve detailed information by clicking on the line and browse to the discussion thread where the matching post is located.

The post count, respectively a single time unit, in the X-axis reflects the occurrence frequency of topic. Additionally information about, diversity of authors who posted that day can be explored. The knowledge about authors diversity can be used to underline for instance how reliable or unreliable is the sentiment in chosen posts. The most important contribution of this visualisation is to offer an overview on diverse discussion forums regard a topic of interest.

5 Preliminary Results

In order to assess our opinion mining system, a baseline-evaluation using two standard information retrieval measures (precision and recall) has been performed. We have compared our approach to a full-text index (Lucene[11]). The domain is currently limited to "cars" (as we have mostly advertisements for the visual analysis available) although we note that the methodology is expected to yield similar results for other domains. The flexibility of our approach is mainly determined by the availability of appropriate instances from DBpedia.

[10] `http://simile.mit.edu/timeplot/`
[11] `http://lucene.apache.org/java/docs/`

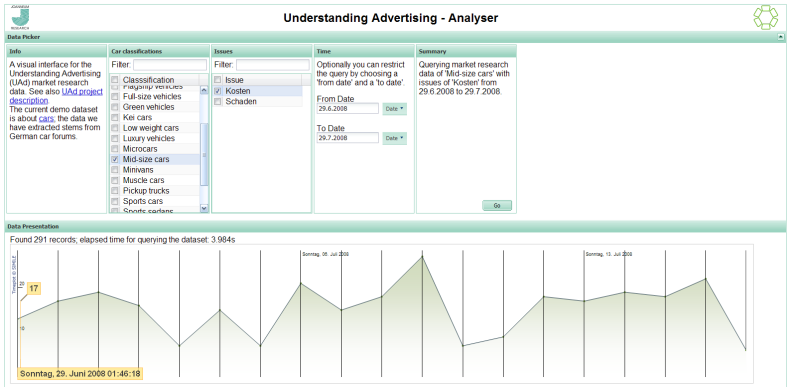

Fig. 5. The UAd Analyser.

5.1 Reference Data Set

Our reference data set contains approximately 1000 posts that have been extracted from a single discussion forum[12], focusing on the content of three sub-forums including all threads and posts about certain car types. Two of the extracted sub-forums contain discussions about cars belonging to the mid-sized car class according to categorisation from DBpedia[13]. The working data set includes 60 representative posts (20 per car type). We have manually selected posts containing discussion on topics such as "performance and problems" and "popularity".

The extracted posts were firstly used to generate opinions on the discussion topics, and secondly for the initialisation of the index over the reference test data (for Lucene). We have converted each of them into a single file containing information on the posting date, author, post URI, the content and the title of the thread the post belongs to, allowing to create an index searchable by Lucence. The Lucene index contains the fields author, title, summary, content and link to post corresponding to the properties in RDFised data and with the intention to provide as similar as possible initial point to RDFised data, for comparison and measurement of results.

Prior to the manual creation of the triggers for discussion topics we have analysed the initialised fields of the Lucene index for occurrence frequency of specific keywords and the "Zipf" distributions [22]. As depicted in listing 1.1 topic triggers contain words or word stems that serve as annotation events. Opinion generation is initiated by accordance of trigger words with words from the content or title of posts. An example discussion opinion generated in this way is shown in listing 1.2.

[12] `http://www.automotiveforums.com`
[13] `http://dbpedia.org/resource/Category:Mid-size_cars`

```
@prefix : <http://sw.joanneum.at/uad/cars/topics#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix opm: <http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf#> .

:performance_and_problems a opm:Topic;
 dc:subject "performance and problems";
 opm:hasTrigger "damage",
                "performance",
                ...
                "problem" .
```

Listing 1.1. Sample discussion topic snippet.

```
@prefix : <http://sw.joanneum.at/uad/cars/opinions#> .
@prefix utop: <http://sw.joanneum.at/uad/cars/topics#> .
@prefix opm: <http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf#> .

:do11 a opm:DiscussionOpinion;
 opm:about <http://dbpedia.org/resource/Alfa_Romeo_156>;
 opm:in <http://www.automotiveforums.com/vbul...php?t=173469>;
 opm:onTopic utop:performance_and_problems .
```

Listing 1.2. Sample generated discussion opinion.

5.2 Results

For the evaluation we have compared our method with the standard Lucene retrieval results of simple queries. Additionally we had a look at extended Lucene-queries; these extended queries have been used to decrease the influence of a single trigger. Listing 1.3 shows a sample SPARQL query we have used for our approach.

```
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix utop: <http://sw.joanneum.at/uad/cars/topics#> .
prefix opm: <http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf#>

SELECT * FROM <http://sw.joanneum.at/uad>
WHERE {
 ?do a opm:DiscussionOpinion ;
     opm:about ?about;
     opm:in ?in ;
     opm:onTopic utop:performance_and_problems .
 ?about owl:sameAs <http://dbpedia.org/resource/Alfa_Romeo_156> .
}
```

Listing 1.3. Sample opinion mining SPARQL query.

From table 5.2 we learn that regarding recall our method unsurprisingly seems to outperform simple full-text indexing. Even in the extended mode Lucene's precision and recall values are below our approach.

		Lucene		UAd Analyser	
		"performance and problems"	*"popularity"*	*"performance and problems"*	*"popularity"*
Precision	simple	0.4	1	0.76	0.86
	extended	0.2–0.62	0.56–0.86		
Recall	simple	0.1	0.05	0.95	0.6
	extended	0.05–0.8	0.3–0.7		

Table 1. Results from the Evaluation (Lucence vs. UAd).

Although we have used a rather limited working set in this evaluation we are optimistic that the results scale well both regarding size and other domains; further evaluations are in the scope of our current research.

6 Conclusion

In this paper we have proposed a novel approach to opinion mining on the Web by using Web of Data technologies and linked datasets. Our goal is to explicitly model opinions found in discussions on the Web; we have developed an according vocabulary to represent these opinions formally (in RDF) and have reported on an implementation of this approach.

We contemplate on using GoodRelations[14]—an ontology for linking product descriptions and business entities on the Web—in order to more accurately describe the target of an discussion in our realm.

To increase the precision we ponder about extending our opinion mining core mechanism with Natural Language Processing techniques and/or use neural networks to categorise topics automatically. As a part of the sentiment classification we aim to use SentiWordnet [23] or other similar approaches for the creation of opinion ranking based on trigger occurrences and the so called PN-polarity [15] of the content.

Currently, we summarise results visually respectively topics, identities, time and occurrence frequency to mirror the sentiment intention in opinions environment. However, currently we do not dive into sentiment interpretation of opinions. Considering the visual analysis, it is important to mention that sentiment interpretation underlies the judgement of end user and his observation standpoint. Anyway, objective parameters such as time period, identities, number of posts etc. can be evaluated independent of matter of particular interest. For further evaluations, user annotated content like reviews or similar will be used additionally.

[14] `http://www.heppnetz.de/projects/goodrelations/`

[15] P stands for "Positive" and N for "Negative" in this context.

Acknowledgements

The research reported in this paper has been carried out in the "Understanding Advertising" (UAd) project, funded by the Austrian FIT-IT Programme. The authors would like to thank their colleagues Magdalena Lauber, Wolfgang Weiss, and Werner Bailer for their support and valuable comments.

References

1. P. Mika. Microsearch: An Interface for Semantic Search. In *Proc. of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008) , Tenerife, Spain*, volume 334 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.
2. W. Halb, Y. Raimond, and M. Hausenblas. Building Linked Data For Both Humans and Machines. In *WWW 2008 Workshop: Linked Data on the Web (LDOW2008)*, Beijing, China, 2008.
3. M. Hausenblas, W. Halb, and Y. Raimond. Scripting User Contributed Interlinking. In *4th Workshop on Scripting for the Semantic Web (SFSW08)*, Tenerife, Spain, 2008.
4. S. Kim and E. Hovy. Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 483–490, Morristown, NJ, USA, 2006. Association for Computational Linguistics.
5. M. Hu and B. Liu. Mining opinion features in customer reviews. In *American Association for Artificial Intelligence at AAAI-04*, 2004.
6. M. Hu and B. Liu. Mining and summarizing customer reviews. In *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining at KDD-2004*, pages 168–177, 2004.
7. K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW2003 - The Twelfth International World Wide Web Conference, Budapest, HUNGARY*, 2003.
8. N. Kobayashi, K. Inui, and Y. Matsumoto. Opinion Mining from Web Documents: Extraction and Structurization. *Informational and Media Technologies 2(1)*, 12(1):326–337, 2007.
9. A. Ghose, P. Ipeirotis, and A. Sundararajan. Opinion Mining using Econometrics: A Case Study on Reputation Systems. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2007.
10. M. Gamon and A. Aue. Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms. In *Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, 2005.
11. M. Hausenblas and H. Rehatschek. mle: Enhancing the Exploration of Mailing List Archives Through Making Semantics Explicit. In *Semantic Web Challenge 2007 at the 6^{th} International Semantic Web Conference (ISWC07)*, Busan, South Korea, 2007.
12. S. Fernandez, D. Berrueta, and J.E. Labra. Mailing Lists Meet The Semantic Web. In *Proc. of the BIS 2007 Workshop on Social Aspects of the Web*, Poznan, Poland, 2007.

13. S. Fernandez, F. Giasson, and K. Idehen. SIOC Ontology: Applications and Implementation Status. `http://www.sioc-project.org/applications#creating-api`, 2007.
14. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, pages 722–735, 2007.
15. J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Inf. Process. Manage.*, 43(4):866–886, 2007.
16. A. Esuli. Opinion Mining. Presentation slides, Language and Intelligence Reading Group, June 14, 2006, Pisa, Italy, Istituto di Scienza e Tecnologie dell' Informazione Consiglio Nazionale delle Ricerche, 2006.
17. B. Liu. Opinion Mining and Summarization, Sentiment Analysis. Presentation slides, Tutorial given at WWW-2008, April 21, 2008 in Beijing, China, Department of Computer Science University of Illinois at Chicago, 2008.
18. T. Berners-Lee. Linked Data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2007.
19. C. Bizer, T. Heath, D. Ayers, and Y. Raimond. Interlinking Open Data on the Web (Poster). In *4th European Semantic Web Conference (ESWC2007)*, pages 802–815, 2007.
20. T. Heath and E. Motta. Revyu.com: a Reviewing and Rating Site for the Web of Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, pages 895–902, 2007.
21. Semantic Web Deployment Working Group. SKOS Simple Knowledge Organization System Reference. W3C Working Draft, Semantic Web Deployment Working Group, 2008.
22. G. K. Zipf. *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology.* Addison-Wesley, 1949.
23. A. Esuli and F. Sebastiani. SentiWordnet: A Publicly Available Lexical Resource for Opinion Mining. In *5th Conference on Language Resources and Evaluation (May 22–28, 2006)*, Genova, Italy, 2006.