Linked Opinions: Describing Sentiments on the Structured Web of Data

Adam Westerski, Carlos A. Iglesias, and Fernando Tapia Rico

Universiadad Politecnica de Madrid,
Escuela Tecnica Superior de Ingenieros de Telecomunicacion,
Avenida Complutense 30, Ciudad Universitaria,
28040 Madrid, Spain
westerski@dit.upm.es,cif@gsi.dit.upm.es,
fernando.tapia.rico@alumnos.upm.es

Abstract. In the paper we report on the results of our experiments on the construction of the opinion ontology. Our aim is to show the benefits of publishing in the open, on the Web, the results of the opinion mining process in a structured form. On the road to achieving this, we attempt to answer the research question to what extent opinion information can be formalized in a unified way. Furthermore, as part of the evaluation, we experiment with the usage of Semantic Web technologies and show particular use cases that support our claims.

Keywords: structured data; ontology; appliance; knowledge management; idea management; opinion mining

1 Introduction

The rise of the Social Web has stimulated progress in many disciplines and gave birth to new trends. One of the research domains that noted especially big progress within recent years is opinion mining. From the information systems point of view, opinion mining aims to harness the flows of unstructured (or poorly structured) subjective user generated textual content that otherwise is hard to analyse, accurately categorise and reason upon. However, while in many cases opinion mining delivers satisfying results it should be aligned to the constantly evolving Web.

One of the problems that we would like to bring to attention is that many web systems (e.g Swotti[1]) that employ opinion mining after gaining understanding of the user generated content, process the extracted parameters (e.g. polarity, features) and publish the outcomes again in an unstructured form (i.e. HTML). On the other hand, others (e.g. Tweet Sentiment[2]) that allow to access the data via web services establish their own formats or languages due to lack of standards that would define clear rules for publishing such information.

[1] http://www.swotti.com/
[2] http://www.tweetsentiments.com/

In our research we aim to show what kind of benefits could it bring to establish a common web metadata schema that would enable to publish in a formalized manner the results of the opinion mining process. As we report on the research done, firstly we introduce the abstract data model - an ontology that formalizes all concepts derived from the opinion mining process (see Sec. 3). Further we propose the use of Semantic Web technologies to adapt that ontology for web use and show exactly what profits can that bring (see Sec. 4). Finally, we present the results of the evaluations run for large scope use cases as well as limited to particular web systems (see Sec. 5).

2 Motivation

Embedding opinion mining functionality for websites that are rich in user comments can aid to automatically rank comments and let users faster reach the types of opinions that they seek [17]. Furthermore, given the same data, opinion mining algorithms can be used to supply additional metrics to rate products and content [20]. However, all of this value is often limited only to the single site of origin that performed the opinion mining algorithm.

Based on the achievements and research done in the area of Semantic Web [7] and more specifically its evolution into proposal of Linked Data [6], we point to publishing opinion information using a universal metadata format that would extend the usability of such data. First and foremost, when having opinions described across the Internet in a unified way it is possible to compare them and perform an Internet wide search and statistics. At the moment it is possible to find opinions of desired polarity about selected product using contemporary Internet search engines, however the simple text based indexing is far less accurate and less flexible than what could be achieved with metadata indexing [11]. Furthermore, if the opinion mining data would be accompanied and linked with other metadata that describes the context of the subjective content, then the capabilities of search and browsing would rise even more (e.g. with regard to aggregation and data mashups [10]).

Finally if all of the above motivations seem fair but far away and hard to realize in practice, we would like to point to what currently seems to be the principal argument for content providers to publish metadata: improve viability on the web and in the search engines. Metadata can help to increase the precision and recall of search [4] but also the value of metadata becomes more visible as the search results in the leading Internet search engines start to contain data extracted form the metadata published along with HTML (e.g. Google Rich Snippets [3]), thus making particular search results more attractive in comparison to competitive links. Through annotation of opinions, exactly the same benefit could be delivered for the websites that provide opinion mining results over subjective content posted on them or remote sources (see Fig. 1).

[3] http://googlewebmastercentral.blogspot.com/2009/05/introducing-rich-snippets.html

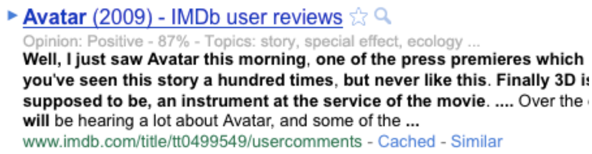

Fig. 1. A Google snippet modified with Greasemonkey script and enriched with data extracted from RDF

3 Marl: An Ontology for Opinion Mining

When designing the ontology our aim was to analyse the properties that characterize opinions expressed on the web or inside various IT systems. The final set of concepts that we propose (see Fig. 2) is a result of a two step process.

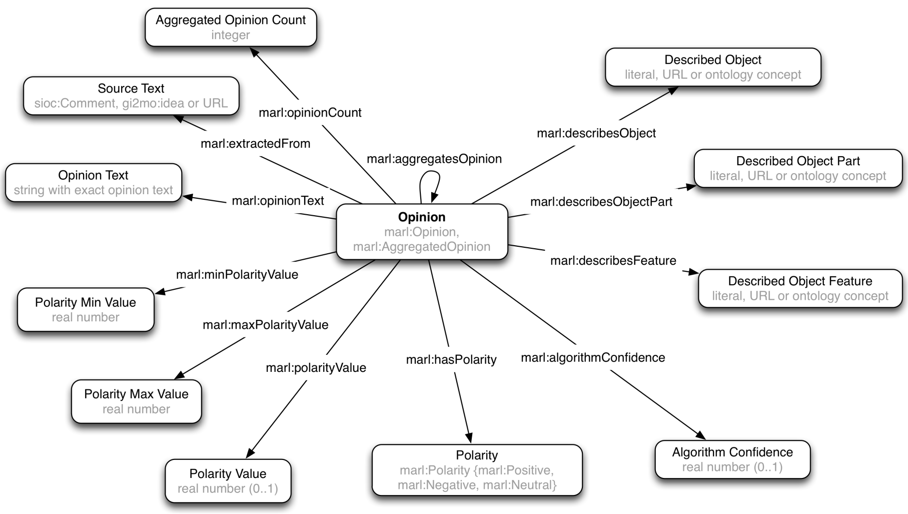

Fig. 2. Conceptual model for opinion and the proposed **Marl** ontology

First, we analysed different kinds of subjective data sources and produced a common model that was formalized as Marl Ontology v0.1. For this part we started with three common cases of opinions expressed on the Web: movie opinions, movie review opinions and products opinions. Later, in addition, we also analysed characteristics of opinions in enclosed communities and used an enterprise open innovation system as a case study. In the second phase, we evaluated the proposed ontology against live data and corrected the discovered drawbacks in version 0.2 of the ontology (see Sec. 5). The description of particular properties and explanation of their meaning can be found in Table 1.

Table 1. Marl ontology: classes and properties breakdown

Name	Description
Opinion	Class that represents the opinion concept
extractedFrom	Indicates the source text from which the opinion has been extracted.
opinionText[1]	The exact string that contains the calculated sentiment.
hasPolarity	Points to either entity or literal that indicates if the opinion is positive/negative or neutral
polarityValue	A numerical representation of the polarity value.
maxPolarityValue	Maximal possible numerical value for the opinion
minPolarityValue	Lowest possible numerical value of the opinion
describesObject	Indicates the object that the opinion refers to
describesObjectPart	Indicates a particular element of the object that the opinion refers to (eg. laptop *battery*)
describesFeature	Points to a feature of an object that the opinion refers to (eg. laptop battery *life*)
algorithmConfidence	A number that describes how much the algorithm was confident with its assessment
AggregatedOpinion	Subclass of Opinion class that aggregates a number of opinions.
aggregatesOpinion	Points to Opinion instances that are aggregated
opinionCount[1]	Amount of opinions aggregated.
Polarity	Instances of this class represent the positive, neutral or negative polarity

In the particular model that we created we attempted to center all the data properties around a single opinion class. This and other ontology design choices that we made with Marl relate to one of the common problems of modelling ontologies for web use: the choice between modelling certain concepts fully as classes of domain ontologies, literals or simply URLs. While for using the full potential of Semantic Web it is best to model metadata concepts as entities described by particular ontologies the reality proves that this is far from being a practical solution. Therefore, we propose a model that accommodates both (see Fig. 3) assuring future extendibility yet facilitating more simple and practical use. In the next section we describe the benefits and applications of either of the cases.

4 Publishing and consuming opinion metadata on the web

Following the description of the opinion ontology we show its possible uses and the differences that various closed and open systems impose. Furthermore, to support the ontology design decisions described earlier, we expose the benefits and drawbacks of publishing opinion data in different forms and with a different level of detail using the Marl ontology.

[1] Properties added in Marl v0.2

```
(1) Example A: Entity referencing for describing contextual information
marl:describesObject <http://dbpedia.org/
                      resource/Avatar_%282009_film%29>
marl:describesObjectPart dbpedia-owl:director
marl:extractedFrom <http://www.gi2mo.org/
                    index.php?sioc_type=comment&sioc_id=157>
marl:polarity marl:Positive
marl:polarityValue "0.6"

(2) Example B: Using literal values to describe contextual information
marl:describesObject "Avatar"
marl:describesObjectPart "director"
marl:extractedFrom <http://www.gi2mo.org/
                    2010/09/introducing-marl/#comment-157>
marl:polarity marl:Positive
marl:polarityValue "0.6"
```

Fig. 3. Referencing entities (1) and using literals (2) with Marl ontology

4.1 Internet wide keyword search and comparison of opinion values

In the simplest case where opinion ontology would be used only with properties expressed with literals, the structure information (connection between opinion text, opinion value and the full body of text) can still be very useful. Even with the contemporary keyword search engines publishing opinion metadata could make a lot of sense. While the discovery of information remains impaired and inaccurate, once actually having found the desired textually expressed opinions, thanks to the metadata it is possible to compare them or transform in different ways. Furthermore, as the research on semantic metadata indexing [15] progresses it is already possible to utilize these simple relationships to make useful search queries on large data sets (see Fig. 4).

```
* <http://purl.org/marl/ns#extractedFrom> * AND
* <http://purl.org/marl/ns#hasPolarity> <http://purl.org/
                                         marl/ns#Positive> AND
* <http://purl.org/marl/ns#describesObject> "Avatar"
```

Fig. 4. Sindice Semantic Index [15] sample query for: "Search positive opinions about Avatar"

4.2 Internet wide entity based search and/or improved data discovery

One of the envisioned bold goals of Semantic Web is to provide entity based search. This would allow to point exact concepts that the user is referring to and eliminate ambiguity of user query present in the keyword search. Slowly this is becoming achievable much due to popularization of big linked data silos (e.g. DBpedia [2]) and wide adaptation of certain ontologies (e.g. GoodRelations [12]). In our research, we also considered using opinion metadata in such scenario. In comparison to the previous case, instead of using literals to describe opinion context Marl ontology properties could point to the exact concepts defined in one of the commonly refereed datasets. This, for example, would allow to formulate queries that distinguish opinions about "Avatar" movie by James Cameron from other meanings of this word (see Fig. 5).

```
* <http://purl.org/marl/ns#extractedFrom> * AND
* <http://purl.org/marl/ns#hasPolarity> <http://purl.org/
                                         marl/ns#Positive> AND
* <http://purl.org/marl/ns#describesObject> <http://dbpedia.org/
                                         resource/Avatar_%282009_film%29>
```

Fig. 5. Sindice Semantic Index [15] sample query for: "Search positive opinions about Avatar" using DBpedia Avatar entity for disambiguation

From a technical point of view, the establishment of such metadata infrastructure would physically link the opinions together with the Linked Data cloud and therefore each other as well via reference to similar topics. In turn, this would allow to traverse the distributed graph in many different ways for numerous use cases, such as aggregation of opinions (see Fig. 6).

4.3 Semantic search engines for dedicated systems

The large scale entity search engines still cope with a number of problems such as insufficient data, efficiency problems etc. even in aforementioned cases of vertical search (e.g. single topic or content type, like the movies). Nevertheless, we also would like to show that similar techniques, that expose the benefits of Marl ontology, can be very useful even if limited to very narrow systems or groups of heterogeneous systems where most of the problems of Internet wide search are eliminated (e.g. in an enterprise).

Following the example of movies that we used in previous cases, the local search could limit to a single website but thanks to the rich data descriptions with the ontological structure it would enable more precise queries than in text search. In this case Marl fills the gap for describing opinions in conjunction with complex taxonomy trees that enable to query for opinions related to particular elements in the class hierarchy that characterizes the given domain.

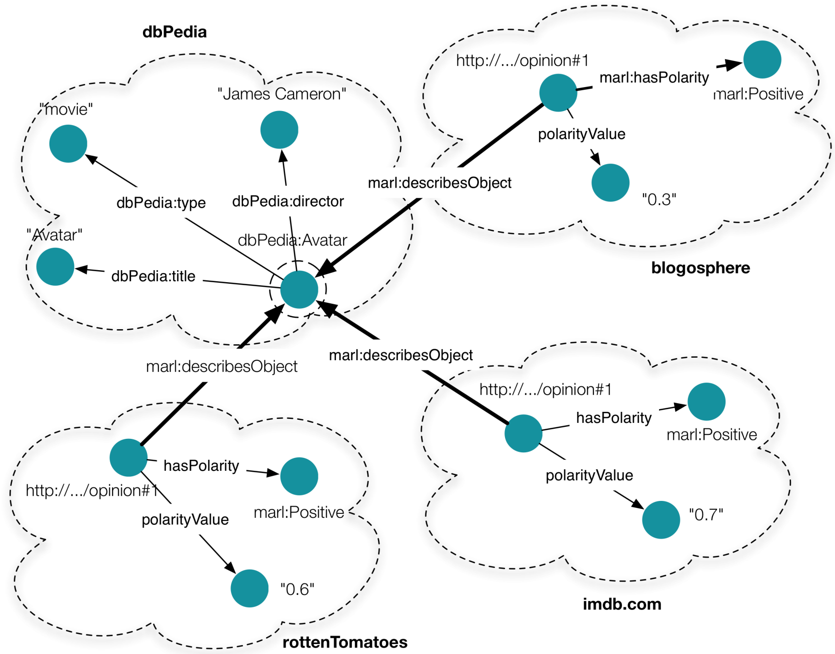

Fig. 6. Sample RDF graph with opinions linked indirectly via metadata references to common entities.

Finally, one can move away from the World Wide Web context to the enterprise environments or other closed systems. In such case the difference is the full control over created data and very strictly defined vocabularies that do not need to be aligned with Web publishing standards. In that case, Marl can be used to together with the verity of enterprise ontologies in the enterprise collaborative systems (e.g. Idea Management Systems or collaborative knowledge management systems). The opinions can be linked via products that they refer to, innovation proposals that are commented by employees, projects in which context the opinions are expressed etc.

5 Evaluation

In order to evaluate our proposal for annotation of opinions we did two experiments. In the first, the goal was to analyse the coverage of the proposed schema against different datasets. In the second experiment we wanted to test in practice how the linked opinion metadata would work with the capabilities of the contemporary search engines and semantic web query endpoints.

During the coverage experiments we analysed two kinds of data: (a) datasets created by other researchers and annotated with opinion mining data; (b) services available on-line that use opinion mining for various goals. The final list consisted of 5 research datasets and 8 services, for each we analysed the data that is exposed and provided Marl mappings. Next, we calculated the coverage as an amount of properties that were possible to describe with Marl over the total amount of data properties used in a dataset. In the first experiment we considered all the dataset fields and the average coverage we got was 64%. However, it has to be noted that the individual characteristics of the data sources varied a lot. According to ontology design goals presented by Noy et al. [14] one of the characteristics of good design is not to cover the very individual elements of datasets. Therefore, after removing the dataset fields that did not repeat at least once, we ran the experiment again and got the average coverage of 76%. The results of the experiments have been summarized in Table 2.

Table 2. Marl ontology coverage experiment results, considering all dataset fields (exp1) and after removing fields that did not repeat at least once (exp2).

Dataset/service name	**#covered/#total**		**coverage**	
	exp1	**exp2**	**exp1**	**exp2**
Congressional speech data [19]	7 / 12	7 / 7	58%	100%
Movie Review Data [16]	3 / 4	3 / 3	75%	100%
Customer Review Data [13]	5 / 9	5 / 6	56%	83%
French Newspaper Articles [8]	1 / 3	1 / 2	33%	50%
Multi-Domain Sentiment Dataset [9]	4 / 4	4 / 4	100%	100%
Swotti (www.swotti.com)	9 / 13	9 / 13	69%	69%
Tweetsentiments (www.tweetsentiments.com)	6 / 11	6 / 11	55%	55%
Mombo (www.mombo.com)	10 / 16	10 / 12	63%	83%
Opinion Crawl (www.opinioncrawl.com)	4 / 9	5 / 9	44%	44%
OPAL (www.gi2mo.org/apps/opal/)	8 / 11	8 / 11	73%	73%
OPfine (www.jane16.com)	6 / 6	6 / 6	100%	100%
Evri (www.evri.com)	3 / 5	3 / 5	60%	60%
Opendover (www.opendover.nl)	4 / 9	4 / 6	44%	67%
Average	**5 / 8**	**5 / 7**	**63%**	**76%**

In the second part of our experiments we tested the capabilities of Marl to be used in context of Semantic Web queries. We started with creating a list of competency questions and tested them against the ontology (a total of 20 query templates where created). Later, for a more practical approach, we extracted small parts of datasets mapped in the previous experiment and used them to check with software prototypes if the queries involving Marl deliver anticipated results with different kinds of search. On this stage the problem that we encountered in most cases was insufficient data to create rich links to expose true power of Marl. Ultimately, for Internet wide data, we did our tests in the context of movie reviews and filtering opinions by polarity from different sites such as Tweetsentiments, IMDB (via Cornell dataset [16]) and Swotti in a single query. We repeated this both for references to movies expressed as literals and for the entity search (with DBpedia entity references). In both cases we used Sindice search engine as back-end for the demonstration. Finally, for tests of metadata search in closed private environments we have setup a local SPARQL endpoint and used the OPAL opinion mining module in conjunction with technologies from Gi2MO project [21] to extract opinions from independent Idea Management Systems and visualise them together. The additional challenge was that the two systems had data in different languages: one Spanish and the other English. As an outcome, the opinion mining algorithm enabled us to leverage the multilingual instances to the same level but ultimately the Marl ontology in conjunction with other Semantic Web vocabularies worked as an enabler to integrate the systems and run queries over the data to aggregate all information in a single view (e.g. show all ideas with community opinions and compare aggregated opinion scores, or compare the amount of positively received ideas by idea categories etc.). All together the query experiments proved that the ontology is capable in answering all the formulated questions in test scenarios of: movie opinions, product opinions, Idea Management Systems. A common problem, that confirmed the test results of coverage experiment, was that many queries expected the direct link to text fragment of the opinion - which is not facilitated by Marl. An example of a query constructed for data serialized with Marl v0.1 during our experiments can be seen at figure 7.

```
PREFIX gi2mo: <http://purl.org/gi2mo/ns#>
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX marl: <http://purl.org/marl/ns#>
SELECT ?idea_uri
       COUNT(?negative_opinion_uri) AS ?negative_opinions
       COUNT(?positive_opinion_uri) AS ?positive_opinions
FROM <http://etsit.gi2mo.org/etsit_ideas_en.rdf>
WHERE {
{
    ?idea_uri a gi2mo:Idea .
    ?idea_uri gi2mo:hasComment ?comment_uri .
            ?positive_opinion_uri marl:extractedFrom ?comment_uri .
            ?positive_opinion_uri marl:hasPolarity marl:Positive .
}
UNION {
    ?idea_uri a gi2mo:Idea .
    ?idea_uri gi2mo:hasComment ?comment_uri .
            ?negative_opinion_uri marl:extractedFrom ?comment_uri .
            ?negative_opinion_uri marl:hasPolarity marl:Negative .
} } GROUP BY ?idea_uri
```

Fig. 7. A sample SPARQL query for "Show amount of positive and negative opinions for all ideas submitted into the Idea Management System". The source data was serialized using Marl v0.1 therefore aggregation operator was used to go around the lack of opinion count information.

Concluding both of the experiments, we used the acquired knowledge to produce a second iteration of the ontology (Marl 0.2) and included the new properties that according to our tests were uncovered and often used by other datasets; or were expected as output for search queries. After repeating the coverage experiments with the new version of the ontology we got 79% coverage for experiment 1 (all dataset fields considered) and 94% coverage for experiment 2 (dataset fields that did not repeat at least one time across different sources ignored).

6 Related Work

The research presented in this paper is primary focused on developing a universal model for describing and comparing opinions on the World Wide Web. As such, it is tied to efforts of the Semantic Web research community, which goals have been outlined by Sir Tim Berners-Lee [5]. Furthermore, as much as we are interested in reasoning and giving birth to the intelligent web, our research is focused to a much more extent on the sole goal of publishing and consuming data. Therefore, we have aligned our investigation with the efforts undertaken by the Linking Open Data project[4] - an attempt to build an interlinked Web of Data using Semantic Web technologies.

In terms of related research conducted in those areas, to our knowledge, there has been only one attempt to achieve a similar goal as our. Softic et al. [18] has proposed an opinion ontology and performed a number of experiments to show its use. However, as authors claim themselves the ontology is unfinished and missing the key element of opinion formalization leaving it for later research which has not done yet. In our work we aimed to use the opinion mining as a tool in our main research area of Innovation Management, therefore we needed a full solution for metadata publishing that could be applied in practice. In comparison to Softic et al. we propose a different conceptual model for the opinion ontology, deliver new properties that describe not only a generic concept but enable to publish the numerical values from the opinion mining process (which is impossible using Softic et al. opinion ontology). Furthermore, with our research we propose a different evaluation framework and test our solution in different cases, which in the end delivers new conclusions and opens new possibilities (see Sec. 7).

Within commercial services related to the area of opinion mining there are different data serialization methods used for APIs but all use own vocabularies. In relation to our work, a standing out service by Opendover moves towards the Semantic Web technologies but the vocabularies used refer only to individual sentiments (thus being more similar to a dictionary) rather than full opinions like in case of Marl ontology.

On the other hand, not related to opinion mining, we recognize that for a practical solution, opinions could be conceptually modelled as reviews. Therefore, in terms of related work we also considered vocabularies created for describing online reviews. Among those, the most popular are: hReviews [1], the RDF

[4] http://esw.w3.org/SweoIG/TaskForces/CommunityProjects /LinkingOpenData

mapping of hReview [3], Google's RDF vocabulary for reviews[5] and Schema.org Review vocabulary[6]. In comparison to our work the existing review formalization vocabularies are much more generic and conceptually describe less referring to the entire review body, whereas we see that the opinion ontology needs to describe particular elements of the review and features discussed in the review (e.g. one might imagine a query using both concepts "show all sci-fi movie **reviews** that contain positive **opinions** about director"). Furthermore, we see reviews as judgement based on factual information and comprehensive knowledge whereas opinions are less formal, smaller pieces of information. For those reasons we believe there is a need for making a distinction between the two concepts in terms of metadata and web search.

7 Conclusions and Future Work

In the paper we have presented a solution for describing opinions on the web with well known and widespread metadata standards of Semantic Web. Furthermore, we have shown how adapting the available metadata specification can help to link opinions with other concepts on the web and lead to better search capabilities and improved exposure of data. Whereas, the full potential of the solution depends on the adoption of W3C recommendations such as RDF or RDFa, we have proven that even with the minimal use of entity search, the publishing of metadata about opinions can be very beneficial. In terms of future work, our aim is very much related to more specific domain research and usage of the Marl ontology in synergy with dedicated ontologies to provide complex search facilities in enclosed systems, very much in a manner as described in the article when referring to vertical search engines and search engines for dedicated systems.

Acknowledgements

This research has been partly funded by the Spanish Ministry of Industry, Tourism and Trade through the project RESULTA (TSI-020301-2009-31) and Spanish CENIT project THOFU. We express our gratitude to Paradigma and Atos Origin R&D for their support and assistance as well as providing us access to their software.

References

1. Allsopp, J.: Review and resume microformats: hreview and hresume. In: Microformats: Empowering Your Markup for Web 2.0 (2007)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: In 6th Int'l Semantic Web Conference, Busan, Korea (2007)

[5] http://www.google.com/support/webmasters/bin/answer.py?answer=146645
[6] http://schema.org/Review

3. Ayers, D., Heath, T.: RDF Review Vocabulary Specification. http://hyperdata.org/xmlns/rev/ (2007)
4. Beall, J.: How google uses metadata to improve search results. The Serials Librarian 59(1) (2010)
5. Berners-Lee, T.: Semantic web road map. http://www.w3.org/DesignIssues/ Semantic.html (1998)
6. Berners-Lee, T.: Linked data. http://www.w3.org/DesignIssues/LinkedData (2006)
7. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American pp. 29–37 (May 2001)
8. Bestgen, Y., Fairon, C., Kerves, L.: Un baromètre affectif effectif: Corpus de référence et méthode pour déterminer la valence affective de phrases. In: Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (2004)
9. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the Association for Computational Linguistics (ACL) (2007)
10. Catasta, M., Cyganiak, R., Tummarello, G.: Towards ecsse: live web of data search and integration. In: Semantic Search 2009 Workshop (SemSearch2009) (2009)
11. Hemminger, B.M., Saelim, B., Sullivan, P.F., Vision, T.J.: Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts. J. Am. Soc. Inf. Sci. Technol. 58(14) (2007)
12. Hepp, M.: Goodrelations: An ontology for describing products and services offers on the web. In: Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW2008). vol. 5268, pp. 332–347. Springer LNCS, Acitrezza, Italy (2008)
13. Hu, M., Liu, B.: Mining and summerizing customer reviews. In: Proceeding of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2004)
14. Noy, N.F., McGuinness, D.L.: Ontology development 101: A guide to creating your first ontology. Tech. rep., Stanford Medical Informatics, Stanford (2001)
15. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Tummarello, G.: Sindice.com: A document-oriented lookup index for open linked data. International Journal of Metadata, Semantics and Ontologies (2008)
16. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL (2004)
17. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2) (2008)
18. Softic, S., Hausenblas, M.: Towards opinion mining through tracing discussions on the web. In: Social Data on the Web (SDoW 2008) Workshop at the 7th International Semantic Web Conference (2008)
19. Thomas, M., Pang, B., Lee, L.: Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: Proceedings of EMNLP (2006)
20. Tian, P., Liu, Y., Liu, M., Zhu, S.: Research of product ranking technology based on opinion mining. In: Second International Conference on Intelligent Computation Technology and Automation (2009)
21. Westerski, A., Iglesias, C.A., Rico, F.T.: A model for integration and interlinking of idea management. In: Metadata and Semantic Research: 4th International Conference, MTSR 2010. Alcalá de Henares, Spain (2010)