Inverting Semantic Structure Under Open Domain Opinion Mining

Boris Galitsky[1], Josep Lluis de la Rosa[2] and Gábor Dobrocsi[3]

[1] Univ. Girona Spain
bgalitsky@hotmail.com

[2] EASY Innovation Center, Univ Girona Spain
peplluis@eia.udg.edu

[3] Univ Miskolc Miskolc Hungary
gadomail@gmail.com

Abstract

We explore the semantic structure of how opinions on products and services are expressed in blogs and forums in the form of user needs. To optimize the efficiency of content delivery, we invert the product feature structure and propose a specific way to represent the user opinion content in forums and blogs, focusing on user needs about product qualities and features. The content is subject to inversion so that these needs become primary entry points for browsing and search. *User need* is defined syntactically; semantic and concept structure means for such *user needs* are developed. The system is subject to evaluation with respect to coverage and information access efficiency.

Introduction

In recent years, blogs and forums became an important source of information about products and services, where experts share their experience with beginner users. Making a buying decision, most users consult forums for opinions, browsing existing forum postings and starting new forum threads is becoming an essential decision support mechanism (Lawrence & Pennock 2003). However, it is quite hard to find a relevant forum posting, or, starting a new one, to receive a prompt and comprehensive recommendation. The reasons for difficulties of relevant information access in such noisy sources as forums and blogs, while making buying decisions are as follows:

1) Distributed nature of blogs and forums   hard to find the one which contains information matching current user interests and needs. To form an opinion about a product feature, multiple sources have to be consulted. It is hard to find a resource to get an immediate response for a posting.
2) Limited trust to particular sources of information and lack of ways to rate authors.
3) Substantial difficulties in indexing blog and forum content for search.

In case of forums, supporting search relevancy by machine learning of which hits have been selected by users, is not very helpful since most likely a local maxima of the relevancy of accepted document will be achieved. Hence deeper understanding of natural language forum postings is required, as well as a new way to represent forum content on what people like and dislike about products and services, and what are their needs and concerns.

The paper proposes processing distributed semantic structure of opinions about products to improve *access efficiency, relevancy and trustworthiness* of opinion data, in a domain-independent manner. We aim at processing blog and forum data to optimize the ease of accessibility for product recommendation with focus on *user needs* about product usability rather than just product features. We propose the grouping of forum content based on products (which is traditional, see Sista & Srinivasan, 2004; Popescu & Etzioni 2005) and then grouping based on natural language expressions of what users like and dislike about products (which requires a specific semantic technology). As a result, we represent blogs, *inverting the content* based on user sentiments, so the user can find features of products based on her needs directly, and proceed towards *associated needs* when necessary. Inversion of blog content therefore allows addressing user needs irrespectively of order, associated discourse of forum postings, and specifics of argumentation patterns. It is expected to be a more uniform, coherent and relevant way of content delivery.

Extracting user need for product recommendation

A vast number of linguistic and statistical studies explored the structure and strength of sentiments, including (Kim & Hovy 2004). In this study we focus on such linguistic structures as *needs* which occur in sentences *under the scope of sentiment*. This class of linguistic structures is an extension of what is traditionally referred to as *features/topics* in literature on opinion mining, towards a general notion of *user needs* and product usability.

In this paper we are concerned with accurate extraction and aggregation of *individual user need expressions* (review quote). When purchasing online, consumers are interested in researching the product or service they are looking to purchase. Currently, this means reading through reviews written on websites of different vendors that happen to offer the product or service. For example, if the consumer is interested in purchasing a digital camera, several on-line vendors allow consumers to post reviews of cameras on the website. Gathering information from such reviews is a discouraging process nowadays as there is little way to sort the reviews for the features that are of interest to any one potential buyer so the potential buyer must read through them manually. Sometimes reviewers rate a product with a given number of stars in addition to making comments.

An average high or low number of stars is not necessarily very informative to a potential buyer, especially if he or she is especially concerned about certain features on the camera. For example, a potential buyer may want a camera from which the photographs come out with very true colors as opposed to oversaturated colors. Other features, such as the weight of the camera or the complexity of the controls are of lesser concern to this potential buyer. A review with many stars may explain the ease of changing batteries of the camera and a review of few stars may complain that the camera only has a 3X optical zoom. Neither of these reviews is relevant to the potential buyer. In order to determine that however the potential buyer must wade through comments, if any, provided by the reviewer that explain why the reviewer scored the. We need a method which would allow for, in addition to the overall sentiment and topic determination, local extraction to determine a specific quote from the analyzed forums and blogs that include the sentiment about that topic. Such a quote would be essential for product recommendation, to serve as an argument for why this particular product is recommended.

Inversion of content

In this section we define the inversion of content for a blog, forum, or aggregated collection of opinions for a product. In various sources of opinions, in different postings about an entity such as digital camera, we combine the textual expressions about a particular need of product users *Subject,* occurring with various parameters in *ParameterList1*. Our goal is to automatically represent the content in the way grouped by Needs, where a content entry will be *Need* with *ParameterList*. The intended transformation can be illustrated in Fig. 1.

Posting1 Forum1 : Sentiment 1 Need(Subject, ParameterList1)
Posting1 Forum2 : Sentiment 2 Need(Subject, ParameterList2)
Posting2 Forum1 : Sentiment 3 Need(Subject, ParameterList3)
Posting2 Forum2 : Sentiment 4 Need(Subject, ParameterList4)

Subject ParameterList1 : Posting1 Forum1
ParameterList2 : Posting1 Forum2
ParameterList3: Posting2 Forum1
ParameterList4: Posting2 Forum2

Fig. 1**:** Illustration for the inversion of content based on user needs.

We now illustrate inversion of content taking into account posting by authors $a \in A$ about needs $f \in F$ of products $p \in P$, Fig. 2.

A typical posting is a request to share information, response to such request or opinion sharing without request, mentioning how is the author related to product domain, whether he likes / dislikes the product itself, its parameter, feature or a particular need, and usability for particular purpose:

Responding to a request:
I am a *beginner* user of a digital camera.
I enjoyed its *zoom* because *it allows taking shots of mountains*.
I used it for *outdoor*

Notice that all italicized expressions are user needs associated with particular product, including product features and their usability. We use graphs to represent in which form this kind of information is available to readers of blogs and forums.

On the left of Fig. 2, the *original* graph for information distributed through blogs and forums is shown. From right to left, authors (nodes a_1, a_2 and a_3) are sharing their opinions on products (nodes $p_1 \ldots p_5$). Each 'opinion sharing' arc is associated with a posting above and is labeled with the content of opinion, a few need expression from the set $\{f_1 \ldots f_6\}$. Needs occur in labels of arcs under positive (f_1) or negative ($-f_1$) sentiment. $\{f_1 \ldots f_6\}$ are *raw* features as expressed by authors. These features are obtained by extracting need expression from text by finding their boundaries; no modification/rephrasing is applied. This original graph reflects the original semantic structure of information submitted by various authors with different product needs and various reputations.

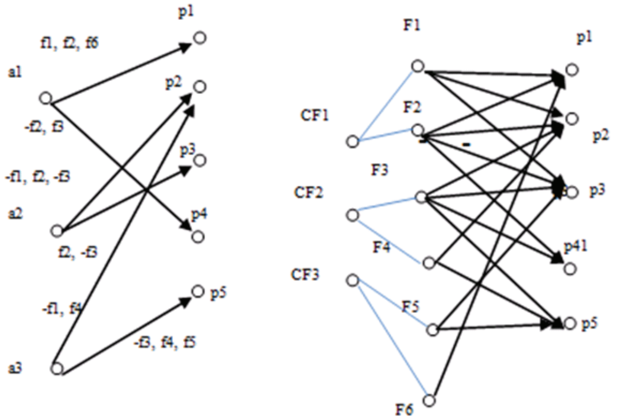


Fig. 2: The graph for original and user need-centric inverted content.

On the right of Fig. 2, the graph for the *inverted* semantic structure of forum and blog data is shown. This is a user need centric representation, where information is 'digested' and converted into a form ready to 'consume' opinions. The inverted graph have the same set of nodes for products $p_1 \ldots p_5$. Now the fact the product p has a feature or need F is expressed by the arc (F, p) of the right graph. Under the process of content inversion, F is a *derived* from the raw feature f from the original (left) graph by a series of transformations described in the rest of this paper, including rephrasing of natural language expression, extraction linguistic patterns, grouping similar needs, finding consistent set of needs and others. Hence the mapping from the original graph to inverted graph converts a-nodes with f-labels into new F-nodes from the derivatives of f. F-nodes are constructed in a way allowing *categorization* of needs.

Extracting user needs from text

In study we are interested in how users express their needs about products and services, so we can extract the needs as natural language expressions and then in a formalized way, suitable for grouping. User 'needs' are semantic structures, but we need a set of syntactic constraints to be applied to text for the purpose of extraction (Galitsky 2003).

These syntactic constraints turn out to be *attachment to a sentiment expression*. To express them, we need to use syntactic tree, where both vertices (lemmas) and edges (syntactic links) are labeled. In a sentence, we first identify sentiment as a vertex (single word like 'good', or subtree 'did not work for me') and then proceed to the *sub-tree which is dependent* (linked to) the main vertex in sentiment sub-tree. Over the years, we accumulated our own domain-independent vocabulary of English sentiments, coded as parsing sub-trees to be identified at parsing trees (compare with Sista & Srinivasan 2004).

Let us consider the domain of digital cameras, and focus on a particular class of usability needs associated with taking pictures at night. We use a pair of tags: *night* + specific *night-related* need sub-category:

night picture (general, overall taking pictures at night)
night>cloud (how to film clouds at night),
night>cold (how to film at night in cold conditions
night>recommend (which measures are recommended at night, general issues)
night>dark (filming in dark conditions)
night>set (what and how needs to be set)
night>inconsistent (for some cameras, setting seemed inconsistent to some users)
night>shot (peculiarities about night shot)
night>tripod (use of tripod at night)
night>mode(switch to specific filming modes for night shots)

As one can see, the meanings for needs of filming at night vary in generality and semantic roles, and phrasings include nouns, adjectives and verbs. So the criteria of being a user need indeed have to be formulated in terms of a sub-tree, satisfying certain syntactic (tree) conditions (see Galitsky 2006 for more details).

For a horizontal (unlimited) domain (like electronics, which is rather wide), all terms from need expressions cannot be defined in an ontology. Therefore, semantics of a need expression has to be inferred from the syntactic one.

Our assumption is that if there is at least one author who attaches sentiment to an expression (which we know now as an expression for need), then other people might have the same need, so it is worth storing and analyzing. In terms of syntactic tree, if a lemma for sentiment is dependent of a term T and does not have its own dependent vertices, the need expression is a sub-tree dependent on T.

Examples of extraction of two need expressions are shown at Fig. 3. For the sentiment 'great', we have a sub-tree 'in-daylight-bright' which is a need expression (use of digital cameras can be 'great', or 'not so good' in 'bright daylight'. For the sentiment '*not*…good', we have a need 'indoor-in-setting-dim. In the latter case sentiment is expressed by 'don't expect it to get good', where the main vertex is 'be', and the need expression is branching from the vertex 'get'.

One needs to differentiate user needs and product features (as presented by manufacturer or retailer). All product features are assumed to be subjects of need, but not otherwise. In terms of natural language, product features and needs are phrased differently. For example, where user need is expressed like 'suited for small fingers', a manufacturer would write '1/4 inch button size'.

Content inversion

After need expressions are extracted, they need to be normalized and grouped. Normalization transforms need expressions into sequences of words in normal form, without prepositions and articles. After that, need expressions are grouped by the main noun of expression (the closest noun to the trunk of the need expression as a sub-tree).

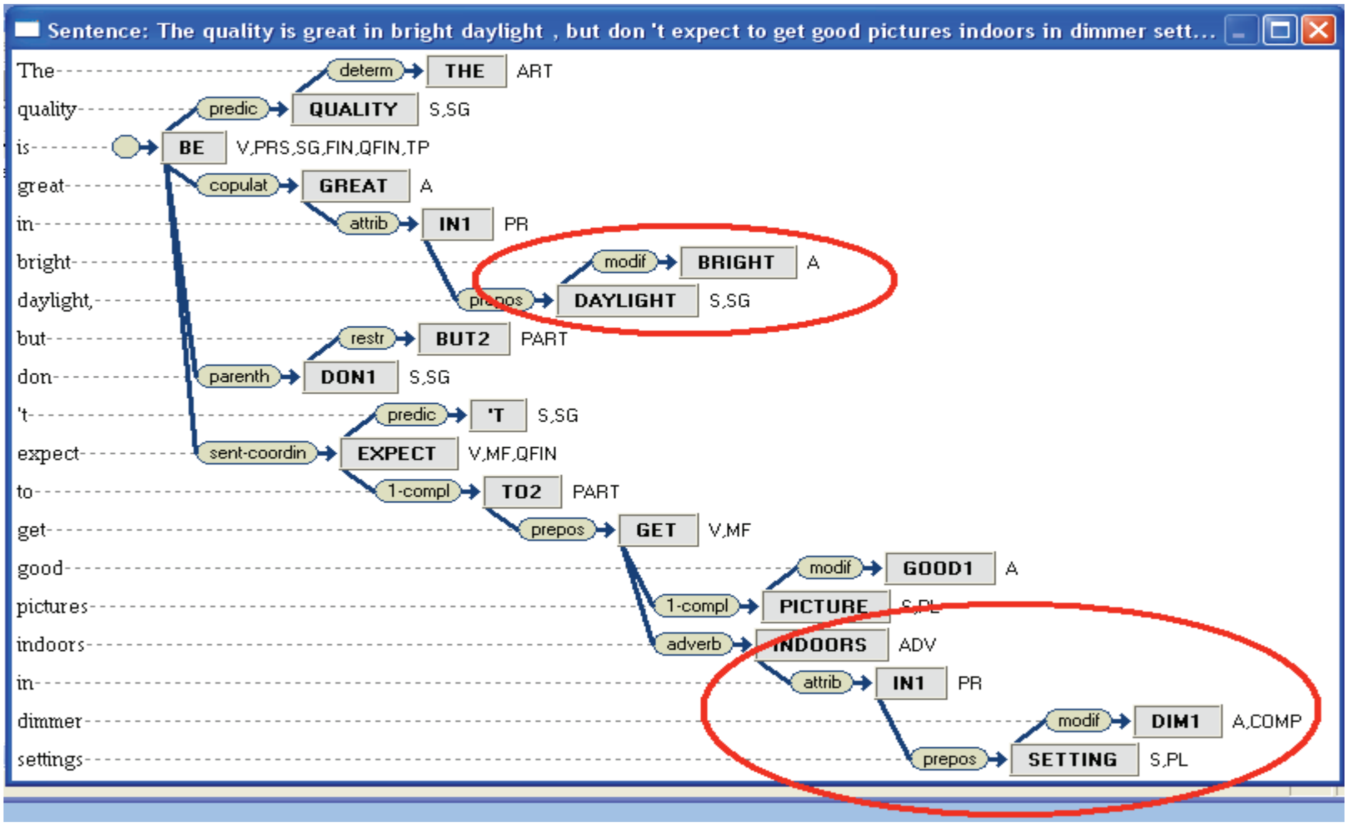

Sentence: The quality is great in bright daylight , but don 't expect to get good pictures indoors in dimmer sett...

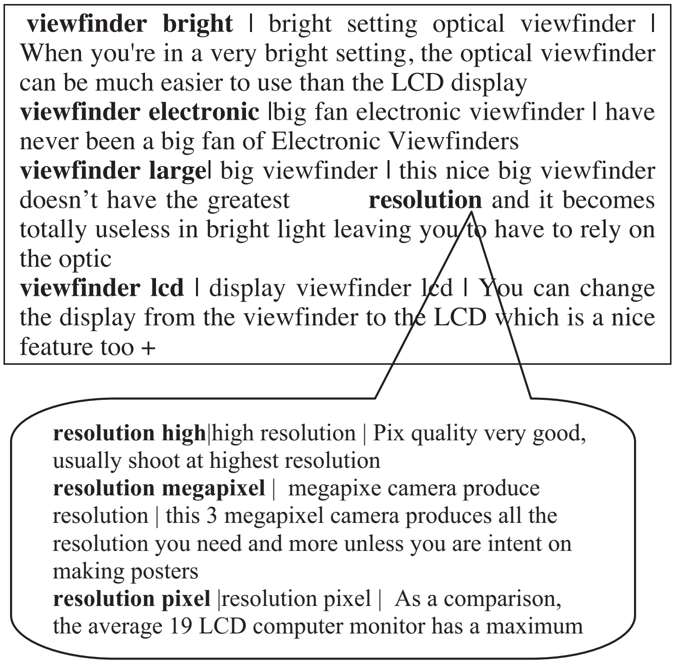
Fig. 3: Syntactic parse tree for sentences about digital camera with two pairs of sentiment-need expressions (circumscribed).

Let us consider an example of a group with noun *viewfinder*, with the second word in grouped expression, all keywords in need expression, and original sentence:

viewfinder>bright | bright setting optical viewfinder | When you're in a very bright setting, the optical viewfinder can be much easier to use than the LCD display
viewfinder>electronic |big fan electronic viewfinder | have never been a big fan of Electronic Viewfinders
viewfinder>large| big viewfinder | this nice big viewfinder doesn't have the greatest resolution and it becomes totally useless in bright light leaving you to have to rely on the optic
viewfinder>lcd | display viewfinder lcd | You can change the display from the viewfinder to the LCD which is a nice feature too

Hence we have four need sub-categories {*bright, electronic, large, lcd*} for the need category *viewfinder*. These subcategories categorize viewfinder from very different aspects. Notice that both syntactic relations between viewfinder and second word varies, as well as semantic relations, however we ignore that for the sake of forming categories and sub-categories.

Four sentences above come from different sources, the common thing between them is the product and a category of user needs about viewfinder in connection to this product.


viewfinder bright | bright setting optical viewfinder | When you're in a very bright setting, the optical viewfinder can be much easier to use than the LCD display
viewfinder electronic |big fan electronic viewfinder | have never been a big fan of Electronic Viewfinders
viewfinder large| big viewfinder | this nice big viewfinder doesn't have the greatest **resolution** and it becomes totally useless in bright light leaving you to have to rely on the optic
viewfinder lcd | display viewfinder lcd | You can change the display from the viewfinder to the LCD which is a nice feature too +

resolution high|high resolution | Pix quality very good, usually shoot at highest resolution
resolution megapixel | megapixe camera produce resolution | this 3 megapixel camera produces all the resolution you need and more unless you are intent on making posters
resolution pixel |resolution pixel | As a comparison, the average 19 LCD computer monitor has a maximum

Fig. 4: Drilling in associated category of needs

Whereas category noun is identified by a rule, a sub-category word is obtained by clustering category into clusters; sub-category word should not be a category word and should occur in more than one need expressions within a category. For more accurate identification of sub-category word more advanced methods could be used, combining machine learning and statistical analysis; it could produce higher percentage of word pairs where meaning can be obtained just from this pair.

Inversion of content is a transformation of corpus of text to a set of interactive textual components where each component includes all content about given need for a given product. These components are hyperlinked to drill in and out of need categories associated with each other.

Let us now draw a hypothetical information access scenario (Fig. 4). If a user is interested in how good is a viewfinder for a given digital camera, all relevant entries are grouped: user can either browse by his need or search by it. Now imagine user got information above, read it and now got interested in 'which viewfinder has a better resolution?'.

When the user (reader) indicates that he is interested in *'viewfinder large'*□ *full sentence* □*'… resolution…',* the system proceeds to the list of needs for the category *'resolution'*. If *'resolution'* is not a category but a sub-category, the system would proceed to the respective sub-category (fewer entries). Otherwise, if 'resolution' occurs in a need expression, such expression will be shown. Finally, if 'resolution' does not occur in any expression, the system retreats to keyword search. This content exploration scenario might be associated with 'hyperlinked text'; in our case hyperlinks and pages are dynamic.

Evaluation

We obtained a few thousand reviews per 100+ digital camera products, built the index for need entries, and provided a basic user interface for browsing and search. The main questions for evaluation are:

1) Coverage: what percentage of user needs can be identified, given the available set of reviews and inversion of this set, implemented in this project (Table 2);
2) Efficiency: how fast (how many steps) it is necessary to find the relevant need entry and get the sentence which describes it (Fig. 5).

Coverage evaluation for 5 queries and averages for another set of 100 queries is shown in Table 1.

To properly interpret accessibility efficiency of the inversion of content, the number of 'steps' should be compared with the number of sentences user would have to read in the body of reviews to reach the sentence which would directly address the user interest. In real life, number of such sentences (including review titles, section titles and directory content) might easily reach 30-50.

To evaluate the *relevancy* of extracted need expression, we built the need-based search framework. In this framework search query is formulated as a certain expression of a user needs about particular features and usability of a product a service, such as 'what kind of cell phone is good for large-size fingers', '… best fits my palm …'. The recall and precision of answers are measured from the standpoint of proper match of needs (Obviously, proper match of need assumes proper identification of products and features themselves). Preliminary estimate of F-measure for such search is above 80% for few product domains processed so far.

Query (expression of interest)	# of need categories explored	# sub-categories explored	Total number of steps	% satisfaction
'Add lens like lens-filter'	2	3	3	80
'Self-timer and flash modes'	4	5	5	60
'Rotate LCD screen in a variety of positions'	2	2	2	80
'Options to increase sharpness and saturation'	2	2	2	100
'Increase ISO to stop camera shake'	1	3	3	70
100 queries (on average)	2.5	3.2	2.8	82%

Table 1: evaluation results

It is quite hard to compare the current results with state-of-art sentiment and topicality extraction because of the way customer needs are defined. Therefore we do not provide the sentiment extraction results here.

Proposed technique has been implemented providing travel recommendations, based on review quotes. In addition to hotel amenities available on current hotel search websites, such need categories as kid-friendliness, pet-friendliness, romance and impressive nightlife are extracted from text and automatically formed (uptake.com).

Conclusions

In this work we performed extraction from text and reasoning about rather general, complex, and abstract object such as user needs about products. This study follows along the line of a body of work about sentiment and polarity analysis (Popescu & Etzioni 2005, Lawrence & Pennock 2003).

The existing research in the area of opinion mining is mainly at the document level, to classify each whole document as positive or negative for particular product feature. In this study, since the user need defined syntactically in terms of a sub-tree, opinion mining can be conducted in an arbitrary domain. To perform the sentence-level inversion of content, sentiments had to be identified individually for proper feature-based grouping of opinions on products. Having defined inversion of content as a formal graph-based transformation, we implemented the systematic and deterministic way of merging similar opinions on the same product features.

We generalized the notion of *feature* extraction towards *need* extraction, which required more sophisticated linguistic analysis means due to significant variability of linguistic parameters for the latter. Feature extraction suffices part-of-speech information, but to circumscribe need expression, full parsing tree is required with detailed labels for nodes and arcs, as well as semantic rules which navigate these trees (Galitsky 2003).

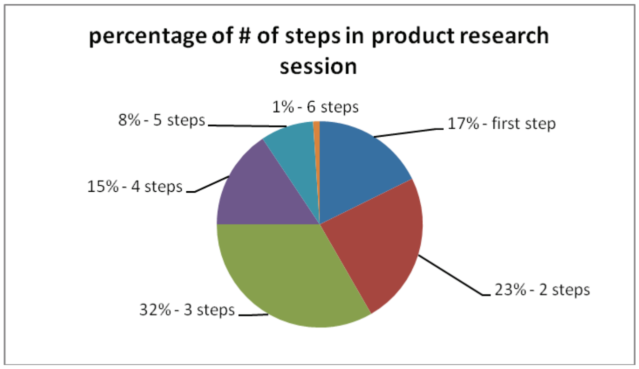


Fig. 5: How fast the user need of interest can be achieved.

Inversion of content can be considered as a semantic-based hypertext generation, quite popular a decade ago (Melucci & Rehder 2003). Automated linking based on lexical and content analysis, which also can be used to determine similarities (relationships) among documents, has been studied. Hypertext functionality includes navigational, annotation, structure-oriented and view-oriented features; however, from the standpoint of given paper automated linking creates static links, unlike the UI presented here.

We observed that inversion of content is an efficient way to access user-generated information in forms of forums and blogs. It is quite obvious that grouping information around entities of interest such as laptops if fruitful for information access and decision support for intended products' features. In this study we proceeded from grouping by entities to grouping and clustering by needs, to accelerate the information access in such area as users' opinions.

Preliminary evaluation showed that proposed approach to semantic-based information access to public opinions provides satisfactory coverage as well as efficient accessibility, compared to conventional browsing and search at social web sites.

Acknowledgements

This research is funded by the European Union project Num. 216746 PReservation Organizations using Tools in AGent Environments (PROTAGE), FP7-2007- ICT Challenge 4: Digital libraries and content, European Union project Num. 238887, *a unique European citizens' attention service* (iSAC6+) IST-PSP, SIE 2007 *Motor intelligent de recuperació i processament semàntic d'informació en base a criteris de personalització*, as well as RDiSAC *Recerca en incentivació de la participació 2.0 i llenguatge natural per contexts per a la creació de serveis no presencials d'informació i atenció ciutadana SAC*, a grant from ACC1O Catalan Government, as well as the CSI-ref.2009SGR-1202 consolidate research group.

References

Popescu, A., Etzioni, O. Extracting Product Features and Opinions from Reviews. Proc. Joint Conf. on Human Lang. Tech. / Conf. on Empirical Methods in Natural Lang. Processing, 339-346 (2005).

Sista, S., Srinivasan, S. Polarized Lexicon for Review Classification. Proc. Intl. Conf. on Machine Learning, Models, Technologies & Applications (2004).

Allen, J. Natural Language Understanding. Benjamin/Cummings (1995).

Galitsky, B. Merging deductive and inductive reasoning for processing textual descriptions of inter-human conflicts. J Intelligent Info Systems, v27, N1, 21-48 (2006).

Galitsky, B. Natural Language Question Answering System: Technique of Semantic Headers. Advanced Knowledge International, Australia (2003).

Melucci, M., Rehder, J. Using Semantic Annotations for Automatic Hypertext Link Generation in Scientific Texts Proceedings of the Workshop on Semantic Web Technologies (2003).

Lawrence, S., Pennock, D. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. Proc. 12th Intl. Conf. on World Wide Web, 519-528. (2003).

Valitutti, A., Strapparava,C. and Stock, O. Developing Affective Lexical Resources. PsychNology Journal, 2(1):61–83 (2004).

Hu, M. and Liu, B. Mining and summarizing customer reviews. In KDD-04, pp 168–177. (2004).

Kim, S.-M. and Hovy, E. Determining the sentiment of opinions. In COLING-2004, pp 1367–1373, Geneva, Switzerland (2004).

Pang, B. and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of ACL-2005 (2005).