Inverting semantic structure of customer opinions expressed in forums and blogs

Boris Galitsky[1] Huanjin Chen[2] and Shaobin Du[3]

[1]Knowledge-Trail Inc. 9 Charles Str. Natick MA bgalitsky@knowledge-trail.com
[2]Uptake, Inc 654 High Str. Palo Alto hchen@uptake.com
[3]Oracle Inc San Mateo CA sdu@oracle.com

Abstract

Abstract: We explore the semantic structure of how opinions on products and services are expressed in blogs and forums. 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 concerns about product qualities and features. The content is subject to inversion so that these concerns become primary entry points for browsing and search. User concern is defined syntactically; semantic and concept structure means for such concerns are developed. The system is subject to preliminary evaluation with respect to coverage, information access efficiency and search accuracy.

Keywords: content inversion, semantic structure, syntactic parse tree.

1 Introduction

In recent years, blogs and forums became an important source of information about products and services, where experts share their experience with beginning users. Making a buying decision, most users consult forums for opinions, browsing existing forum postings and starting new forums is becoming an essential decision support mechanism [7]. 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 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 concerns. 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 around what people like and dislike about products and services.

The paper proposes processing distributed semantic structure of opinions about products to improve *access efficiency, relevancy and trustworthiness* of opinion data,

particularly. We aim at processing blog and forum data to optimize the ease of accessibility for product recommendation with focus on *user concerns* about product usability instead of just product features. We propose the grouping of forum content based on products (which is traditional, see [1,2,11]) 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 concerns directly, and proceed with associated concerns when necessary. Inversion of blog content therefore allows addressing user concerns irrespectively of order, associated discourse of forum postings, and specifics of argumentation patterns, which is expected to be a more uniform, coherent and relevant way of content delivery.

This project was inspired by the idea to combine the social and technological advances of the web infrastructure. The project is expected to leverage combining the features of:

- the social web applications, leveraging network effect and capable of accumulating textual data,
- knowledge representation and reasoning about how features are combined with sentiments,
- linguistic processing with semantic focus,
- and fast-growing online data production via network effects of forums.

Social web infrastructure and the semantic web technologies complement each other in the way they approach obtaining new content and making it accessible. Social web applications are usually trivial at handling semantics of content, providing limited content access capabilities. On the other hand, semantic web applications are better at natural language technologies, but less efficient at user engagement. We expect the content inversion based on user concerns to leverage the best of these two worlds.

A vast number of linguistic and statistical studies explored the structure and strength of sentiments, including [12, 13]. In this study we focus on such linguistic structures as *concerns* 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* in literature on opinion mining towards a general notion of *user needs* and product usability.

2 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 cameras, we combine the textual expressions about a particular concern of product users *Subject,* occurring with various parameters in *ParameterList1.*Our goal is to automatically represent the content in the way grouped by Concerns, where a content entry will be *Concern* with *ParameterList.* The intended Transformation can be depicted at chart Fig. 1.

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

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

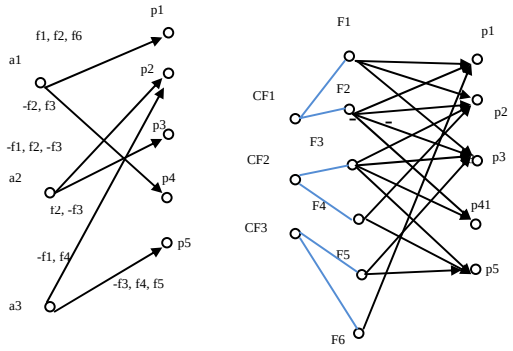
Fig. 1. illustration for the inversion of content based on user concerns.

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

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 concern, 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 concerns 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, 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 concern expression from the set $\{f_1 \ldots f_6\}$.

Concerns 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 concern 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 [8].

On the right, the graph for the *inverted* semantic structure of forum and blog data is shown. This is a product/feature – 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 concern 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 concerns, finding consistent set of concerns 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 concerns.

This conversion takes into account inconsistencies in the opinion of conflicting authors. If two authors express concerns about the same feature of the same product of the opposite polarity, we try to resolve the conflict by determining the *reputation* of authors and their competence relevant to given concerns. For example, reputation of an author is higher if he has a higher number of postings about the relevant subjects. Conflict can be solved in favor of the author with significantly higher reputation; if it is impossible to resolve, F-feature is not formed.

To automatically perform the inversion of content, the following problems need to be solved:

1) Defining concerns as syntactic structures to extract from text;
2) Finding boundaries of concern expression based on syntactic parse tree.
3) Defining inversion as a graph transformation and implementation of such transformation;
4) Building semantic model to group extracted concerns;
5) Filter out irrelevant and inconsistent concerns by inductive learning;
6) Visualize products and concerns for interactive exploration (by a concept lattice);
7) Matching user search query with concern expression.

In this article we briefly outline 1) -4) & 6); problem 5) requires sophisticated machine learning of syntactic graphs and is a subject of our further studies. Problem 7) is a subject of a separate study as well [4].

The proposed process relies on the following assumptions:

1) Each concern is expressed within a single sentence, we don't need to analyze sentence co-references [3] to extract concerns;
2) Sentiments are independent of concerns.

The possibility of inversion of content is based on our assumption that there is a one-to-many mapping between sentences and concerns expressed in these sentences, and the majority of sentences in reviews make sense being stand-alone (and attached to the entry in inversed content).

3 Extracting user concerns from text

In study we are interested in how users express their concerns about products and services, so we can extract the concerns as natural language expressions and then in a formalized way, suitable for grouping. User 'concerns' are semantic structures, but we need a set of syntactic constraints to be applied to text for the purpose of extraction [5].

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 [2]).

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

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 concerns of filming at night vary in generality and semantic roles, and phrasings include nouns, adjectives and verbs. So the criteria of being a user concern indeed have to be formulated in terms of a sub-tree, satisfying certain syntactic (tree) conditions (see [4] for more details).

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

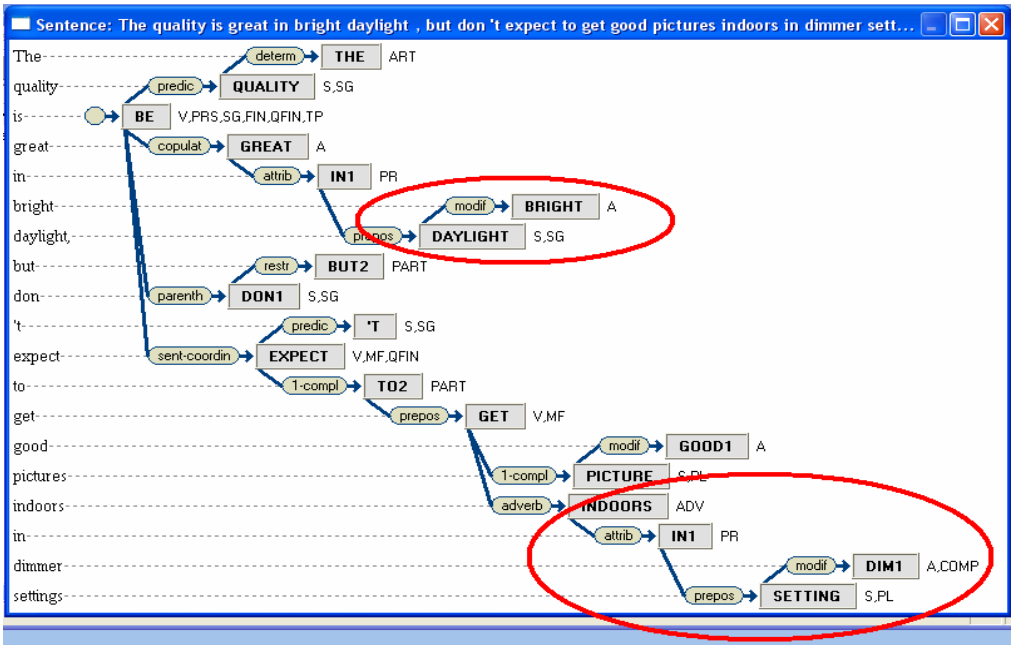


Fig. 2. Parse tree for sentences about digital camera with two pairs of sentiment-concern expressions (circumscribed).

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 concern), then other people might have the same concern, 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 concern expression is a sub-tree dependent on T.

Examples of extraction of two concern expressions are shown at Fig. 2. For the sentiment ‘great’, we have a sub-tree ‘in-daylight-bright’ which is a concern expression (use of digital cameras can be ‘great’, or ‘not so good’ in ‘bright daylight’. For the sentiment ‘*not…*good’, we have a concern ‘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 concern expression is branching from the vertex ‘get’.

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

4 Content inversion

After concern expressions are extracted, they need to be normalized and grouped. Normalization transforms concern expressions into sequences of words in normal form, without prepositions and articles. After that, concern expressions are grouped by the main

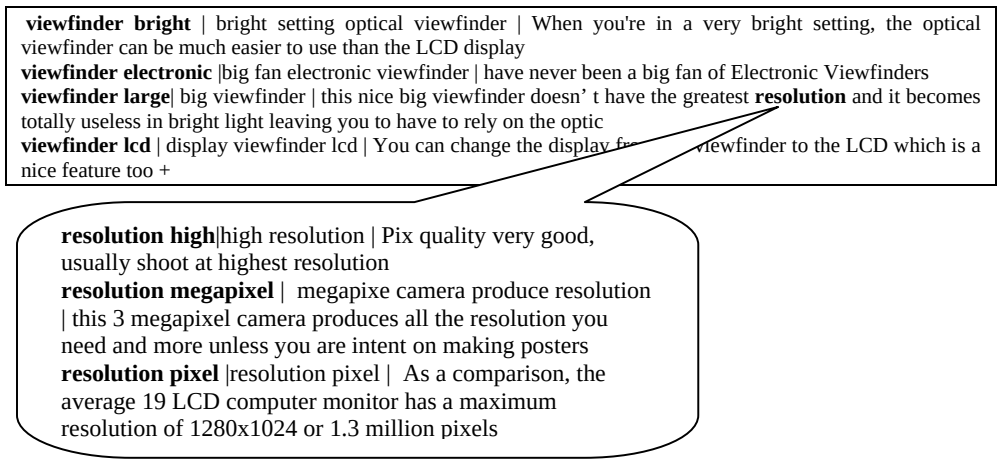
noun of expression (the closest noun to the trunk of the concern expression as a sub-tree).

Let us consider an example of a group with noun *viewfinder*, with the second word in grouped expression, all keywords in concern 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 concern sub-categories {*bright, electronic, large, lcd*} for the concern 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 concerns about viewfinder in connection to this product.



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 concern 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 components where

each component includes all content about given concern for a given product.
Let us now draw a hypothetical information access scenario. 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 concern 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 concerns 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 concern 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 and search-based.

5 System Architecture

System architecture of inversion of content – based forum is shown in Table 1. On the left, offline preparation steps are shown. Notice that two rounds of linguistic processing are required: first round prepares concern expressions for grouping and filtering, and the second round prepares concern entries. Off-line component build the index for search as well, so that user query or short posting or message can be matched against one or multiple concern entries.

On the right, the online functionality is shown. Both 'search' and 'browsing' for relevant concerns are supported in a homogeneous manner, based on the indexed concern entries. Having chosen the domain, user can drill-in to more specific concern, explore concerns at the same level or move up for a new set of concern entries.

When similar concerns about the same products are grouped, best products for given concerns, and most important concerns for products can be explored using concept structures (Fig3; we use visualization by ConExp [9]). Concern data are exported into the ConExp system automatically.

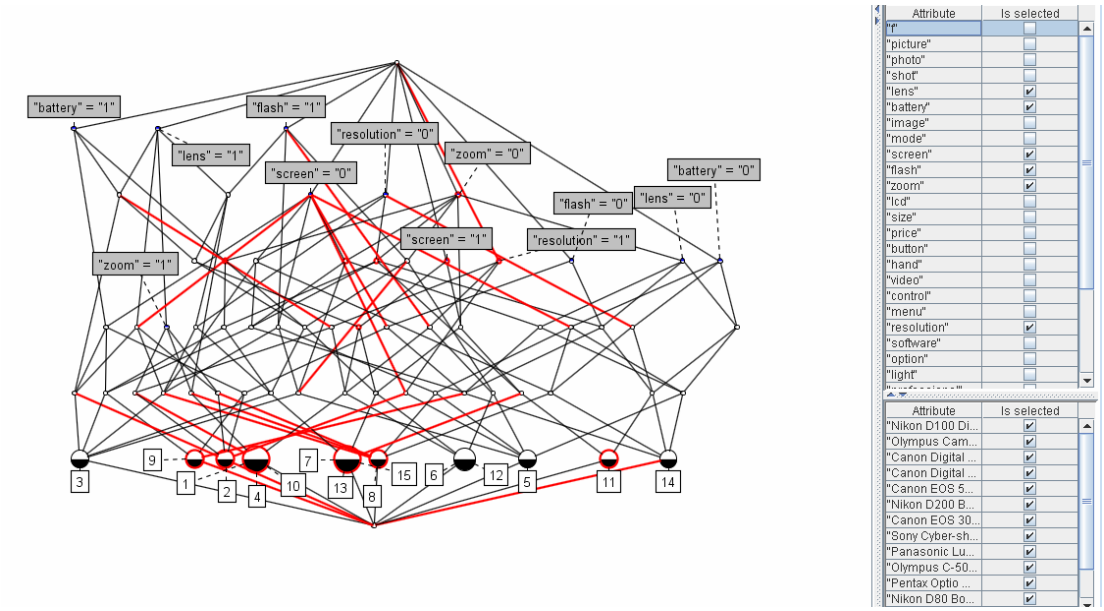


Fig.3. Screen-shot of interactive exploration of product' features, extracted from blogs and forums

6 Preliminary evaluation

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

1) Coverage: what percentage of user concerns can be identified, given the available set of reviews and inversion of this set, implemented in this project;
2) Efficiency: how fast (how many steps) it is necessary to find the relevant concern entry and get the sentence which describes it (Fig. 4).

Coverage evaluation for 5 queries and averages for another set of 100 queries is shown in Table 2.

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 concern expression, we built the concern-based search framework. In this framework search query is formulated as a certain expression of a user concerns 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 concerns (Obviously, proper match of concern 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.

Table 1. System architecture

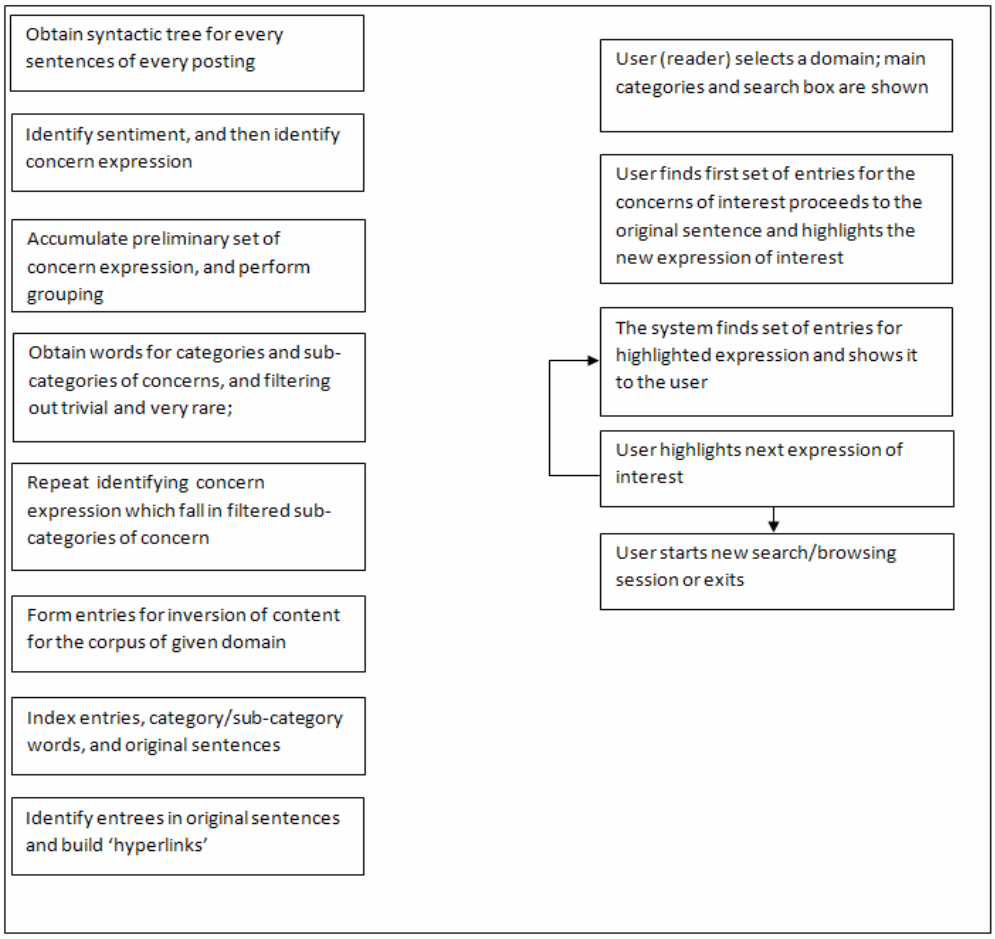


Table 2. Evaluation of coverage for concern representation.

Query (expression of interest)	# of concern categories explored	# sub-categories explored	Total number of steps	% satisfaction
'Add lens like circular polarizer lens'	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%

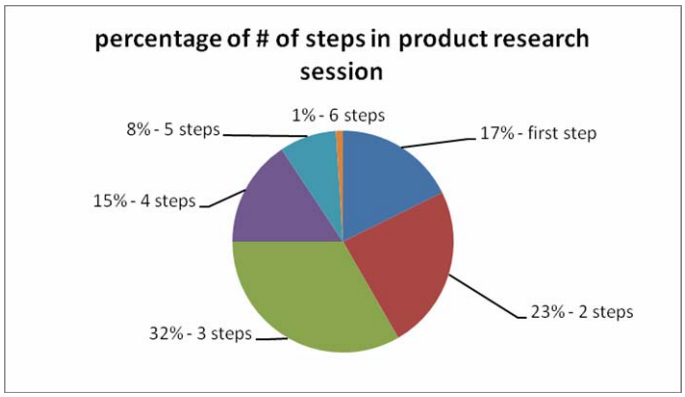


Fig. 4. How fast the concern of interest can be achieved, using concern-based inversion of content.

7 Conclusions

In this work we performed extraction from text and reasoning about rather general, complex, and abstract object such as user concerns about products. This study follows along the line of a body of work about sentiment and polarity analysis [1,7].

The existing research in the area of opinion mining is mainly at the document level, to classify each whole document as positive or negative (we assume neutral belongs to positive). To perform the sentence-level inversion of content, sentiments had to be identified individually for proper feature-based grouping of opinions on products. We generalize the notion of *feature* extraction towards *concern* 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 concern expression, full parsing tree is required with detailed labels for nodes and arcs, as well as semantic rules which navigate these trees [5].

This project can be considered in the framework of semantic-based hypertext generation, quite popular a decade ago [6,14, 15]. 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 made the next step, proceeding from grouping by entities to grouping and clustering by concerns, to accelerate the information access in such unstructured 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.

References

1. 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).
2. Sista, S., Srinivasan, S. Polarized Lexicon for Review Classification. Proc. Intl. Conf. on Machine Learning, Models, Technologies & Applications (2004).
3. Allen, J. Natural Language Understanding. Benjamin/Cummings (1995).
4. Galitsky, B. Merging deductive and inductive reasoning for processing textual descriptions of inter-human conflicts. J Intelligent Info Systems, v27, N1, 21-48 (2006).
5. Galitsky, B. Natural Language Question Answering System: Technique of Semantic Headers. Advanced Knowledge International, Australia (2003).
6. Melucci, M., Rehder, J. Using Semantic Annotations for Automatic Hypertext Link Generation in Scientific Texts Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data (2003).
7. Dave, K., 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).

8. Galitsky, B and Levene, M. Simulating the Conflict Between Reputation and Profitability for Online Rating Portals. Journal of Artificial Societies and Social Simulation vol. 8, no. 2 http://jasss.soc.surrey.ac.uk/8/2/6.html (2005).
9. Yevtushenko, S.A. http://www.sf.net/projects/conexp. Last accessed Jan 7 2009.
10. Valitutti, A., Strapparava,C. and Stock, O. Developing Affective Lexical Resources. PsychNology Journal, 2(1):61–83 (2004).
11. Hu, M. and Liu, B. Mining and summarizing customer reviews. In KDD-04, pp 168–177. (2004).
12. Kim, S.-M. and Hovy, E. Determining the sentiment of opinions. In COLING-2004, pp 1367–1373, Geneva, Switzerland (2004).
13. Pang, B. and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of ACL-2005 (2005).
14. Bieber,M., Kacmar, C. Designing hypertext support for computational applications Communications of the ACM (1995).
15. Allan, J. Automatic hypertext link typing. In Proceedings of the Seventh ACM Conference on Hypertext (1996).
16. Riloff, E., Wiebe, J. and Wilson, T. Learning subjective nouns using extraction pattern bootstrapping. In *Conf. on Natural Language Learning (CoNLL)*, pages 25.32 (2003).