Ontology-based corpus generation for web comment analysis

M. Neunerdt, T. Cury Teixeira,
R. Mathar,
Institute for Theoretical Information Technology
52074 Aachen, Germany
{neunerdt,teixeira,mathar}@ti.rwth-aachen.de

B. Trevisan, E. Jakobs
Textlinguistics/ Technical Communication
52062 Aachen, Germany
{b.trevisan,e.m.jakobs}@tk.rwth-aachen.de

ABSTRACT

In the study of technology acceptance the survey and analysis of user opinions is crucial to identify acceptance-relevant factors. In addition to topic-related surveys, which require a high user willingness, the World Wide Web poses a huge collection of user discussions and comments to serve as a basis for further analysis. An ontology-based corpus generation tool is proposed, which serves to evaluate and extract particular web comments.

General Terms

Algorithm

Keywords

Ontology modeling, blogs, user behavior, Web 2.0

1. MOTIVATION

Web 2.0 applications are no longer limited to passive viewing of website content, but allow for user interactions and collaboration in a virtual community. Along with an increasing number of Web 2.0 sites, most of the discussions about the risk of large-scale technologies take place online. Typical examples are social networks, blogs and forums. People post articles or blog posts, which are in turn commented and evaluated by readers. Opinion and market research institutions already make use of such data in order to require opinions about products and services. For this purpose, blog comments are collected and analysed with the goal of identifying relevant evaluation factors.

The aim of this work is to adapt this approach to investigate the acceptance of mobile communication systems (MCS). The goal is to identify key factors inhibiting and promoting the acceptance of mobile communication systems and detect their dependencies on each other.

For this purpose, an ontology-based corpus generation tool is developed, which at the same time serves for ontology evaluation based on a given web comment data collection. It allows for extracting particular web comments by means of a predefined ontology. This paper introduces the tool with its current functions and components.

2. THE CORPUS GENERATION TOOL

The increasing number of user generated content allows for accessing large amounts of web data. However, since people think in different ways and use different terminologies for discussions it becomes hard to access and extract topic-specific user opinions.

The tool, named CROW, aims at selectively searching for web comments that are relevant to a predefined set of topics for further analysis, respectively. In this approach not only topic-specific terms are used but also background knowledge containing concepts represented by terms and their semantic relations; they form the ontology and serve to evaluate the relevance of web comments.

Particularly, in the field of Knowledge Engineering and Semantic Web Research ontolgies are of strong interest. Typical applications are, i.e., mapping between concept hierarchies (ontology mapping) [1] and ontology-based focused crawling [5, 2]. Ontologies contain a collection of concepts which exist or may exist in a certain domain and their semantic relations among each other. The relations are determined according to linguistic usage of terms or respectively to human semantic associations. Thus, the ontology can be thought of as a directed graph where the nodes represent terms and the edges represent semantic relations. The edges describe the dependency between MCS components, properties and instances. Hence they serve to focus on web comments dealing with a particular subtopic in the context of MCS. In total, six relations are distinguished [3]: *Hyperonymy:* The hyperonymy is a hierarchical relation as it can be found in taxonomies. Several subordinated concepts(hyponyms) are assigned to a superordinated concept (hypernym). Hyperonymy appear in the ontology as a unidirectional relation. The edges are labeled with 'IS-A'. *Synonymy:* The synonymy is a semantic relation that is based on identity or similarity. In the ontology, synonyms are summarized into a concept or term set (synset). Synonyms in the synset are interconnected via undirected edges. The edges are labeled with 'SAME-AS'. *Meronymy:* The meronymy describes a part-whole relation. It indicates a constituent part of another concept. In the ontology it is displayed as a unidirectional relation and labeled with 'PART-OF'. *Property:* Property relation indicates object properties (e.g., robustness). The edge appears unidirectional in the ontology.

It is labeled with 'HAS-PROPERTY'. *Association:* The association is a relation type that is neither conceptual nor lexical in nature. Concepts are related to each other on the basis of experience. The relation between two associated words is labeled with 'SEE-ALSO'. *Instance:* The instance assigns real world examples to ontology concepts or nodes (e.g. iphone as an example of smartphone). The edge is unidirectional aligned and labeled with 'INSTANCES'.

In this approach an ontology for the MCS terminology is created and analyzed to identify domain-specific terms. By means of a frequency analysis for a given web comment data collection, see Section 4, the terms are identified. Focus of the analysis is the identification of nouns, opinionated adjectives and characteristic noun phrases.

3. TOOL FUNCTIONALITIES

CROW gives the possibility to manually enter the ontology or choose an existing one. Nodes can be added or deleted. Each ontology can be saved at any time for future work. Various statistic scores, like term frequencies, the total number of comments, the number of comments including single terms/combination of related terms, etc. are determined and plotted in the graphical user interface to aid the user to evaluate and extract relevant web comments. Particularly, the correlation coefficient for related terms is calculated. We consider two connected nodes n_i, n_j with corresponding term frequency vectors $\mathbf{F_i} = (f_{i1}, \ldots, f_{iU})^T$, $\mathbf{F_j} = (f_{j1}, \ldots, f_{jU})^T \in \mathbb{R}^U_+$, where U is the number of comments. Based on these frequency vectors the weight is calculated as

$$w_{ij} = \frac{\sum_{u=1}^{U}(f_{iu} - \bar{f}_i)(f_{ju} - \bar{f}_j)}{\sqrt{\sum_{u=1}^{U}(f_{iu} - \bar{f}_i)^2 \sum_{u=1}^{U}(f_{ju} - \bar{f}_j)^2}},$$

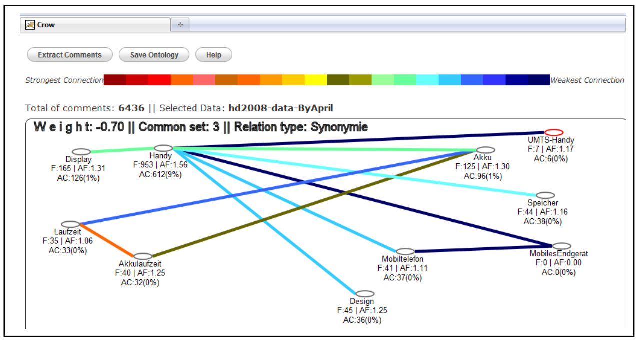
where $\bar{f}_i$ and $\bar{f}_j$ are the frequency means of $\mathbf{F_i}$ and $\mathbf{F_j}$. The correlation strength between those frequency vectors is depicted by the colour of the relation line with a given colour scale. Users can interpret and use those relation weights to choose a particular ontology path for comment extraction. For instance, calculating the correlation allows for interpretation how often a synonym is used. Future work intends to identify and calculate particular weights for different relation types considering, e.g., part of speech and word distance, which indicate the relation type itself.

4. EXEMPLARY EVALUATION

A set of weblog pages from 2008 and 2009 is collected. Applying some preprocessing steps, posted comments with their corresponding meta information, e.g., the posting date, the posting time and the users name are extracted. Afterwards the files are separated into different folders according to users name and time stamp, respectively. Table 1 contains some statistics about the corpus. Particular user comments and, e.g., monthly comments are selected as data corpus in CROW for further linguistic analysis, [4]. Exemplarily, Figure 1 shows a CROW screenshot representing the results based on comments of April 2008 for a MSC subontology.

5. OUTLOOK

CROW allows for generating different web comment corpora concerning a predefined ontology for certain users or times. The generated corpora serve to identify the key factors influencing the acceptance of mobile communication systems in further work.



Total of comments: **6436** || Selected Data: **hd2008-data-ByApril**

Weight: -0.70 || Common set: 3 || Relation type: Synonymie

Figure 1: Screenshot CROW.

Table 1: Data Statistics

	Data 2008	Data 2009
Articles	1252	1289
Comments	84203	81831
Users	10474	9509
Mean number of comments		
per article	67	63
per user	8	9

Furthermore, the approach aims at analyzing different relation statistics based on the current web comment data collection. The goal is to determine weights, which help to detect a particular type of semantic relation. By means of those weights, new frequently used terms can be added with the correct semantic relation to the currently existing ontology. Hence, CROW cannot only be used for corpus generation but also for automatic ontology generation and expansion.

6. ACKNOWLEDGEMENT

This work was partially supported by the Project House HumTec at RWTH Aachen University, Germany.

7. REFERENCES

[1] Z. Aleksovski, M. C. A. Klein, W. ten Kate, and F. van Harmelen. Matching unstructured vocabularies using a background ontology. In *European Knowledge Acquisition Workshops*, 2006.

[2] M. Ehrig and A. Maedche. Ontology-focused crawling of web documents. In *Proceedings of the ACM Symposium on Applied computing.*

[3] S. Roman. Eine Ontologie für die Grammatik. Modellierung und Einsatzgebiete domänenspezifischer Wissensstrukturen. In *Konferenz zur Verarbeitung natürlicher Systeme*, 2006.

[4] B. Trevisan and E.-M. Jakobs. Talking about mobile communication systems: Verbal comments in the web as a source for acceptance research in large-scale technologies. In *Professional Communication Conference*, 2010.

[5] H.-T. Zheng, B.-Y. Kang, and H.-G. Kim. An ontology-based approach to learnable focused crawling. *Information Sciences*, (23), 2008.