# Deriving Case Base Vocabulary from Web Community Data

Kerstin Bach, Christian Severin Sauer, and Klaus-Dieter Althoff

University of Hildesheim
Institute of Computer Science - Intelligent Information Systems Lab
Marienburger Platz 22, 31141 Hildesheim, Germany
{bach,althoff}@iis.uni-hildesheim.de
christiansauer@gmail.com
http://www.iis.uni-hildesheim.de

**Abstract.** This paper presents and approach for knowledge extraction for Case-Based Reasoning systems. The recent development of the WWW, especially the Web 2.0, shows that many successful applications are web based. Moreover, the Web 2.0 offers many experiences and our approach uses those experiences to fill the knowledge containers. We are especially focusing on vocabulary knowledge and are using forum posts to create domain-dependent taxonomies that can be directly used in Case-Based Reasoning systems. This paper introduces the applied knowledge extraction process based on the KDD process and explains its application on a web forum for travelers.

**Key words:** case-based reasoning, knowledge container, vocabulary extraction

## 1 Introduction

Since the very beginning and the expansion of the World Wide Web (WWW) the information available increased rapidly. This information overhead lead to new technologies and developments that ended up in the Web 2.0. Moreover, these technologies enhanced the information palette for multimedia data and boosted the amount of user-generated content. Further, the Web 2.0 changed the documents provided in the WWW from rather statical documents on a website towards smaller pieces of user-driven information like forum posts, tweets[1] or ratings for products [19]. The user generated content often contains pieces of user experiences expressed in an explicit or implicit way[15]. However, using these information is still challenging because most of it is still unsystematic and therewith hardly to be efficiently retrieved and reused [3]. In the last years many kinds of web communities that had some kind of structure to the experience available were mostly develop around a certain type of topic and assemble WWW users that share an interest. Nevertheless the amount of web communities that develop based on common interests is still quite large, however, those

[1] So called 140 character messages on twitter.com

communities can be classified into a manageable amount of community types. Those community types can be characterized by their technical realization and the way how knowledge and experiences are presented and shared. Both aspects can be used to differentiate between communities and provide a first type of structure [12]. The information provided in web communities can be described as a combination of experiences. So the information web have to deal with contains common information and information artifacts like web sites or documents that are interwoven with information given during interactions or conversations in web communities.

### 1.1 The Role of Knowledge Extraction within SEASALT

The work presented in this paper is a part of the further development of the SEASALT (Sharing Experience using an Agent-based System Architecture LayouT) architecture [16], which is especially suited for the acquisition, handling and provision of experiential knowledge as it is provided by communities of practice and represented within Web 2.0 platforms. It is based on the Collaborative Multi-Expert-Systems (CoMES) approach presented by Althoff et. al. [1] which is a continuation of combining established techniques and the application of the product line concept (known from software engineering) creating knowledge lines [10]. The knowledge extraction is a part of the knowledge provision component in SEASALT that is based on software agents representing heterogeneous domains. Each software agent can be realized as an independent Case-Based Reasoning system that is maintained by a *Case Factory* [2]. Each agent retrieves new knowledge from individual knowledge sources like wikis, blogs or web forums in which users provide different kinds of information.

### 1.2 Discussion of Related Work

According to Richter [17], the knowledge of Case-Based Reasoning systems can be provided in all four knowledge containers that include vocabulary, similarity measures, transformational knowledge and cases. We focus on the problem (task) on how the knowledge containers can be filled using the experiences provided in web communities. For this purpose we have to extract the knowledge before it can be stored.

Further on, we deal with data sources that are mostly free text what usually requires a symbolic representation of keywords. That is the reason why we currently focus on the extraction of taxonomies that can be used for both, enhancing the vocabulary and assigning the similarity. Since we use user experiences to develop our taxonomies, we are able to create taxonomies tailored to represent experiences provided in the according community. However, besides extracting cases from experiences there are many more pieces of knowledge that can be extracted and used by Case-Based Reasoning systems. The work of Smyth et. al. [19], Milne et.al [14], Plaza and Baccigalupo [15] as well as Ihle et. al [9] show the potential of web communities for Case-Based Reasoning systems.

For the generation and the improvement of a Case-based Reasoning system's vocabulary or its domain model different knowledge extraction approaches can be applied. For example, light ontologies can be used for domain modeling and directly extracted from the web community. This can be carried out by creating taxonomies after analyzing social tags, like the extraction of folksonomies [13]. The benefit of using light ontologies and folksonomies is quite clear: the more people are working and contributing in a web community, the more detailed the result gets with almost no increase of the cost.

Another major point while extracting vocabulary knowledge is the fact that the information has to be put into a certain context to be organized in a knowledge model. Mika [13], Chakrabarti [4] and Liu [11] for example present approaches in which they take the role of the user into account. Even if only little information is available: certain properties of the information, the explicit or implicit scope of the web community, or the role of a user can provide information on the context.

## 2 Knowledge Extraction

Knowledge acquisition is still the bottleneck within the development of Case-Based Reasoning systems. Our approach aims at automatization and for this purpose we propose a schema that extracts knowledge from web community sources and provides it for CBR systems. In the remaining parts of this paper we introduce our approach and apply it in our current real-life application docQuery. docQuery is a SEASALT instantiation that provides travel medicine information based on data extracted from a web community [16].

Currently we work on the extraction of taxonomies that are used to build the vocabulary, to assign similarity values between two symbols as well as provide basic adaptation knowledge. The vocabulary is derived by transferring each term of a taxonomy into the vocabulary repository. Taxonomically ordered symbolic types provide similarity values between two nodes based on their distance in the taxonomy and sibling nodes in taxonomies represent basic adaption possibilities.

Besides the creation of taxonomies, also the improvement of those taxonomies, the adjustments of term weights for the similarity assignment as well as the possibility to resolve disambiguation also has to be focused on. The aforementioned approaches take the web usage or the links pointing at a certain resource of the extracted terms into consideration. Techniques from Information Retrieval and the analysis of social networks are necessary to put the extracted information in the domain context. The goal of our approach is structuring the available information and organizing them within the knowledge model, so the experiences can used by the Case-Based Reasoning system.

### 2.1 Process

Our approach of extracting knowledge from web communities and applying this knowledge within a Case-Based Reasoning system can be compared to the knowledge discovery in databases (KDD) presented by Fayyard [7].

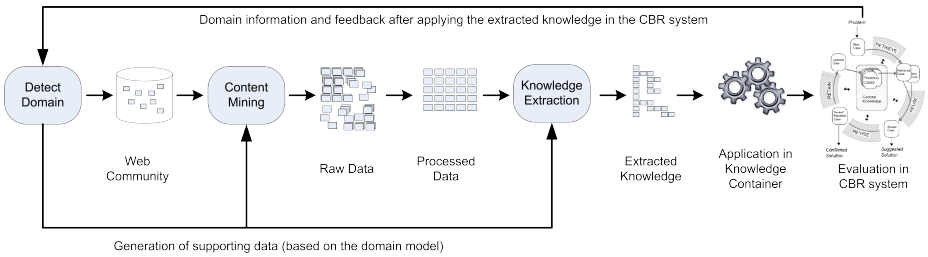


**Fig. 1.** Knowledge Extraction Process Model for CBR Systems

The knowledge extraction process presented here does not only apply for CBR systems, of course it can be employed in all kinds of knowledge based systems. In comparison to KDD, we do not create views on data, instead we are extracting knowledge, reorganize it and provide it in knowledge containers. Therefore we do not concentrate only on one particular data type by applying and refining information extraction techniques on it. Instead, our source data is different, because we get our source data delivered from web mining technologies, most of them are web (content) crawlers. The extraction of knowledge and further on its provision in the Case-Based Reasoning system's knowledge containers can be described as an active transformation of knowledge, because we create and/or extend knowledge models, i.e. taxonomies. Figure 1 gives an overview of the Knowledge Extraction processes based on the KDD process. The following nine steps describe successively what kinds of tasks have to be executed in each step to create knowledge from community data:

1. **Domain Detection:** This first step describes the identification of the domain properties and results in the assignment of what kind of information can be extracted and in which knowledge container it should be integrated. Further auxiliary data like word lists or rules that are required for the knowledge extraction are defined and, if possible, created.
2. **Web Community Selection:** In this step an appropriate web community has to be identified. This covers not only technical issues, it also requires the permission to use the knowledge provided within a web community. The technical requirements include the availability of a certain amount of accessible data. This usually starts with structure mining techniques before continuing with the next step.
3. **Content Mining:** This describes the connection of the knowledge extraction process to the source data. Usually web crawlers access the content and transform it so the data can be provided. Also intelligent web forums, as they are described by Feng [8] and further developed within the SEASALT architecture [16] where intelligent agents pass on relevant data, can serve as source web community.
4. **Processing Raw Data:** Noise, stop words and duplicates are removed and the retrieved data from step 3 is transformed to be processed in the next step. This step ensures that the extraction methods receive "clean" data.

5. **Processed Data:** The result of the refining process applied to the raw data.
6. **Knowledge Extraction:** Based on experience an extraction method (like a Named-Entity-Recognition (NER)-Tagger) has to be selected. This step relies on the methods available and the experiences made in past extraction processes as well as it depends on the kind of knowledge that has to be extracted. Each extraction method requires different types of additional input data like vocabulary lists (stop words), rules or ontologies (WordNet) to carry out the information extraction tasks. In this steps this input is provided (if not already done in the domain model step (step 1)) as well as the information extraction application has to be configured (represented by the lower arrow in Figure 1. Within this step this data has to be provided to finish up the final preparation of the knowledge extraction.
7. **Extracted Knowledge:** The knowledge extraction is carried out by executing the extraction processes using the previously described auxiliary material. After this step the extracted knowledge is ready for further use.
8. **Application in Knowledge Container:** The obtained knowledge is transferred to its according knowledge container so it can be applied in the Case-Based Reasoning process. In case new knowledge models have been built, they have to be integrated in the knowledge representation and made available to the processes they are supporting.
9. **Evaluation:** The knowledge obtained is evaluated using the Case-Based Reasoning system(s): this describes evaluation processes of each of the four kinds of knowledge within its application domain. Information on the quality of knowledge can be obtained to further develop the extraction processes by applying experiences made while executing the knowledge extraction process.

The Knowledge Extraction steps are carried out interactively and the degree of automation highly depends on the web community and the target data. The following sections will introduce an interactive knowledge extraction application that uses a discussion forum to build taxonomies.

## 3 Interactive Knowledge Extraction from Community Data

The knowledge extraction process for Case-Based Reasoning systems we have introduced will now be applied in a real life application domain travel medicine. We use a web forum in which experts, ex-pats and travelers discuss various topics regarding traveling to South East Asia. First, we will introduce the knowledge extraction workbench, the graphical user interface presenting the knowledge extraction results to the knowledge engineers, followed by an explanation of our experimental results after applying the knowledge extraction process on the aforementioned forum.

### 3.1 Knowledge Extraction Workbench

Figure 2 shows the Knowledge Extraction Workbench that we have developed for extracting taxonomies from web community data. This user interface supports the knowledge extraction process steps 4 – 7.

One the left hand side the knowledge engineer can browse forum posts and select the analysis method. The right hand side displays the extracted taxonomies and allows the knowledge engineer to load and save different taxonomies as well as manipulate similarity measures. Further, the knowledge engineer can edit the taxonomy and recalculate the similarity assignments for any node within the taxonomy. A more detailed explanation of the workbench can be found in [18].

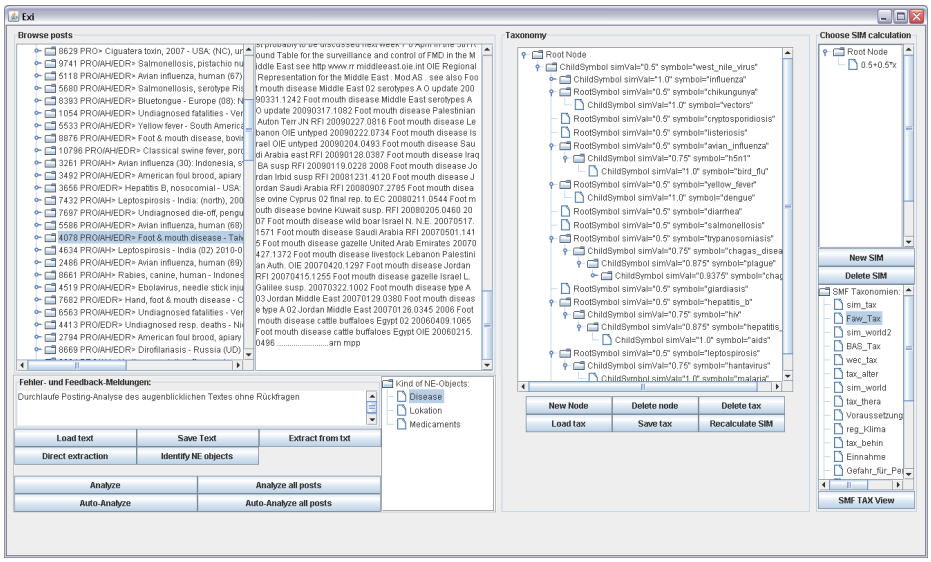


**Fig. 2.** Screenshot of the Knowledge Extraction Workbench

The domain detection and provision of the web community data has been done in advance as well as a web forum equipped with intelligent agents was created. The agents are monitoring the forum and pass on posts that belong to a certain topic, currently we deal with the topics location, diseases and medicaments. In advance, we have ensured that our experts agree with the fact that we further process their posts. The intelligent web forum is built upon a data base so we can access the free text via an odbc connection.

Currently, the workbench supports building taxonomies that can be directly imported in the Protégé[2] plugin of MyCBR [20]. Further on, the workbench is connected to GATE, a text engineering tool [6], that we use for information

[2] `http://protege.stanford.edu/`

extraction. The according GATE component is called ANNIE (A Nearly-New Information Extraction System) that we configured using lists of keywords and rules. The goals of the Knowledge Extraction Workbench is creating taxonomies using forum posts. The forum posts can be analyzed in two ways - either post by post - or the whole thread in the forum. In both cases we do a quantitative analysis of the data and based on the number of occurrences of terms we create a parent node for those terms that occur most frequently and add child nodes that occur second most frequently. We are assuming that according to Church and Hanks [5], two terms are more similar the more often they occur next to each other. For example a forum post contains four times the term *children's disease* and three times the term *measles*. Our approach takes children's disease as parent node and assigns measles as its child node.

### 3.2 Experimental Results

The experiments are carried out on a forum containing posts of travelers, expats and experts who are discussing various issues on travel medicine aspects, especially prevention and travel preparation issues. On the other hand, we have a web forum that contains posts discussing all kinds of travel related topics. In total our experimental data contains about 6500 forum posts and we are extracting knowledge for a medicament, location and disease case base. During the development of the docQuery application, our travel medicine experts provided key word lists for medicaments and diseases that we use in combination with a set of rules for setting up the information extraction environment of ANNIE.

Since the Knowledge Extraction Workbench offers two types of analysis, we tested the performance of those. The first type analyzes the discussion thread in the way that the whole thread is treated as an isolated text. In this process, the term chains are created covering the whole thread. A term chain is a sorted list, descending according to the number of occurrences, of all extracted terms.

In comparison, the second analysis type treads each post as an isolated text and executes the aforementioned process, thus building term chains only covering the isolated post.

Figure 3 depicts two resulting taxonomies with he upper one after the full thread analysis and the lower one post-by-post analysis. The full thread analysis always created deeper taxonomies, because the term chains are longer if built over the whole text of a thread.

## 4 Discussion

The work presented in this paper focused on web communities and how to systematically use the experiences provided by users of this communities in Case-Based Reasoning systems. We are positive that Case-Base Reasoning can handle and benefit from the information given during discussions, because it is a flexible methodology for integrating different kinds of knowledge in a knowledge based

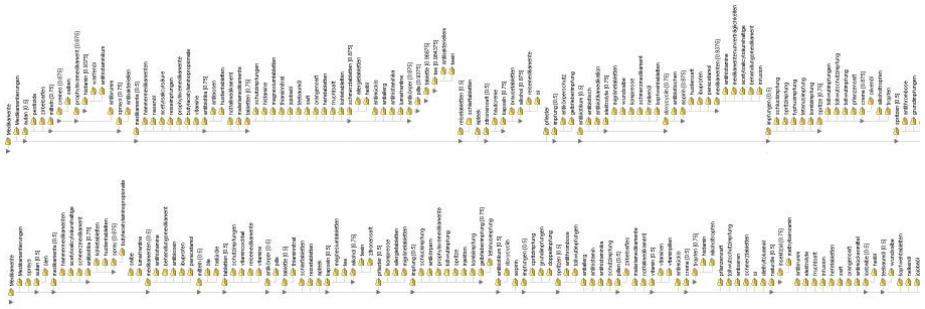

**Fig. 3.** Taxonomy Depth Comparison

system. We have discovered three different scenarios in which the knowledge extraction process we introduced can be applied: First, the knowledge extraction process can be applied once during the development of a Case-Based Reasoning system supporting the system's design. Second, it can be used for initially filling the knowledge containers and third it can be completely integrated in the maintenance process of Case-Based Reasoning systems. Therefore the extraction has to be scheduled regularly and processes have to be defined that allow a further development of the knowledge structure, comprehending updating, evaluating the knowledge and assuring its quality. Within docQuery, we use the last scenario because web communities are the main type of knowledge sources. Using the third scenario of the extraction process allows us to react to changes in these communities and allows us to discover and integrate new terms quickly.

Web Communities already exist on many topics, however, using those for processing user generated content systematically requires the approval of its users. We have used a web forum that we have built for the docQuery project and that has been designed for extracting knowledge from user discussions. Disadvantages of this approach are the requirement of a sufficient amount of users with expertise and the tendency of web based discussions to cover a broad range of topics thus often failing to concentrate on a given topic.

Usually the discussions are centered around a manageable amount of topics. On the other hand, working with data from existing communities requires their permission. We have made the experience that when communicating why and how their data is processed, they usually agree. However, the recent development of the Web 2.0 focuses more and more on the collaboration and participation of users which both provide further support for automatic experience and knowledge processing.

Applying Case-Based Reasoning within the Web 2.0 can also lead towards demand-oriented and web-based systems that integrate their functionalities in daily working routines of users and therewith provide knowledge gained from their contributions and behavior. Therefore, the technologies for knowledge extraction have to be further developed and semantic web technologies, multi-

agent-technologies as well as soft-computing technologies can improve todays systems. Also, a deeper integration of users is possible: it might be feasible to allow the community member actively participate in knowledge modeling.

According to the case factory approach [2], we can also imagine using Case-Based Reasoning systems for supporting or even executing the information extraction. Therefore we have to first fill a case base covering information on what type of extraction method can be applied on which source data to retrieve a certain type of knowledge. In this context, Case-Based Reasoning might be able to keep up with very knowledge intensive IE methods that require additional information like thesauri, gazetteer or annotated training data, because it can handle incomplete information and the similarity-based search might be able to cope with today's problems of achieving an exact analysis and mining of text segments.

## 5 Summary and Outlook

This paper picks up the task of realizing knowledge extraction for Case-Based Reasoning systems from web communities. First we discussed how experiences occur on the web and in which way they can be further used. Afterwards we presented a selection of related approaches before we have introduced a knowledge extraction process that has later on be applied and evaluated in the travel medicine domain. The paper sums up with a discussion of the results and the further potential of the approach we presented.

The knowledge extraction process is a first realization based on well known and applied methods that have been applied in a new way. After establishing a knowledge extraction process we will now concentrate on applying and developing new approaches that improve the quality of the retrieved data or integrate different sources obtained from Web 2.0 developments. For example folksonomies as a result of social tagging analysis can serve as case base vocabulary.

Another aspect of further research is the application of the knowledge extraction process on topics that are more open like free time activities or more general travel experiences. Currently we are focusing on key word centered data and the transition towards the extraction of more various data with more flexible case representations. Further on, the knowledge extraction workbench is still a prototype and has to be improved aiming at a better usability for knowledge engineers.

## References

1. Althoff, K.D., Bach, K., Deutsch, J.O., Hanft, A., Mänz, J., Müller, T., Newo, R., Reichle, M., Schaaf, M., Weis, K.H.: Collaborative multi-expert-systems – realizing knowlegde-product-lines with case factories and distributed learning systems. In: Baumeister, J., Seipel, D. (eds.) Workshop Proc. on the 3rd Workshop on Knowledge Engineering and Software Engineering (KESE 2007). Osnabrück (Sep 2007)

2. Althoff, K.D., Hanft, A., Schaaf, M.: Case factory – maintaining experience to learn. In: Göker, M., Roth-Berghofer, T. (eds.) Proc. 8th European Conf. on Case-Based Reasoning (ECCBR'06), Ölüdeniz/Fethiye, Turkey. Springer, Berlin (2006)
3. Bergmann, R.: Experience Management: Foundations, Development Methodology, and Internet-Based Applications, LNCS, vol. 2432. Springer (2002)
4. Chakrabarti, S., Dom, B.E., Gibson, D., Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Mining the link structure of the world wide web. IEEE Computer 32, 60–67 (1999)
5. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1), 22–29 (1990)
6. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: Gate: A framework and graphical development environment for robust nlp tools and applications. In: Proc. of the 40th Anniv.Meeting of the Assoc. for Comp. Linguistics (ACL'02) (2002)
7. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. Commun. ACM 39(11), 27–34 (1996)
8. Feng, D., Shaw, E., Kim, J., Hovy, E.: An intelligent discussion-bot for answering student queries in threaded discussions. In: IUI '06: Proc. of the 11th Intl Conf. on Intelligent user interfaces. pp. 171–177. ACM Press, New York, NY, USA (2006)
9. Ihle, N., Hanft, A., Althoff, K.D.: Extraction of adaptation knowledge from internet communities. In: Delany, S.J. (ed.) ICCBR 2009 Workshop Proc., Workshop Reasoning from Experiences on the Web. pp. 269–278 (July 2009)
10. van der Linden, F., Schmid, K., Rommes, E.: Software Product Lines in Action - The Best Industrial Practice in Product Line Engineering. Springer, Berlin, Heidelberg, Paris (2007)
11. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2008)
12. Mika, P.: Flink: Semantic web technology for the extraction and analysis of social networks. Journal of Web Semantics 3, 211–223 (2005)
13. Mika, P.: Social Networks and the Semantic Web, vol. 5. Springer Science+Business Media, LLC, Boston, MA (2007)
14. Milne, P., Wiratunga, N., Lothian, R., Song, D.: Reuse of search experience for resource transformation. In: Delany, S.J. (ed.) ICCBR 2009 Workshop Proc., Workshop Reasoning from Experiences on the Web. pp. 45–54 (July 2009)
15. Plaza, E., Baccigalupo, C.: Principle and praxis in the experience web: A case study in social music. In: Delany, S.J. (ed.) ICCBR 2009 Workshop Proc., Workshop Reasoning from Experiences on the Web. pp. 55–63 (July 2009)
16. Reichle, M., Bach, K., Althoff, K.D.: The seasalt architecture and its realization within the docquery project. In: Mertsching, B. (ed.) Proc. of the 32nd Conference on Artificial Intelligence (KI-2009). pp. 556–563. LNCS (Sep 2009)
17. Richter, M.M.: Introduction. In: Lenz, M., Bartsch-Spörl, B., Burkhard, H.D., Wess, S. (eds.) Case-Based Reasoning Technology – From Foundations to Applications. LNAI 1400, Springer-Verlag, Berlin (1998)
18. Sauer, C.S.: Analyse von Webcommunities und Extraktion von Wissen aus Communitydaten fr Case-Based Reasoning Systeme. Master's thesis, Institute of Computer Science, University of Hildesheim (2010)
19. Smyth, B., Champin, P.A., Briggs, P., Coyle, M.: The case-based experience web. In: Delany, S.J. (ed.) ICCBR 2009 Workshop Proc., Workshop Reasoning from Experiences on the Web. pp. 74–82 (July 2009)
20. Stahl, A., Roth-Berghofer, T.R.: Rapid prototyping of cbr applications with the open source tool mycbr. In: ECCBR '08: Proc. of the 9th European conference on Advances in Case-Based Reasoning. pp. 615–629. Springer, Heidelberg (2008)