Model-based Classification of Unstructured Data Sources

Kerstin Bach[1] and Klaus-Dieter Althoff[2]

[1] Verdande Technology AS
Stiklestadveien 1, 7041 Trondheim, Norway
kbach@verdandetechnology.com

[2] University of Hildesheim, Institute of Computer Science
Competence Center Case-Based Reasoning
German Research Center for Artificial Intelligence (DFKI) GmbH
Trippstadter Strasse 122, 67663 Kaiserslautern, Germany
klaus-dieter.althoff@dfki.de

Abstract. In this paper we present an approach that uses knowledge provided in Case-Based Reasoning (CBR) systems for the classification of unknown and unstructured textual data. In the course of developing distributed CBR systems, heterogeneous knowledge sources are mined for populating knowledge containers of various CBR systems. We present how available knowledge, especially the kind of knowledge stored in the vocabulary knowledge container, can be applied for identifying relevant experiences and distributing them among various CBR systems. The work presented is part of the SEASALT architecture that provides a framework for developing distributed, agent-based CBR systems. We focus on the implementation of the knowledge mining task within SEASALT and apply the approach within a travel medicine application domain. Our underlying data source is a user forum, in which various travel medicine topics are discussed, and we show that our approach outperforms the C4.5 and SVM classifiers in terms of accuracy and efficiency in identifying relevant forum entries to create cases from.

Key words: Case-Based Reasoning, Knowledge Mining, Knowledge Containers, Distributed Case-Based Reasoning

1 Introduction

In application domains where heterogeneous data sources contain relevant experiences for Case-Based Reasoning (CBR) systems we are faced with the challenge of identifying, extracting and formalizing such experiences in order to provide them on request. CBR has been proven to provide experiences, however, there is often significant manual effort necessary to collect experiences. In this work, we assume that experiences are cases in a CBR system, which originate in a web forum where users discuss travel medicine topics. These topics usually cover

among others the target region along with disease, medicament, activity and/or environmental information. Our goal in the work presented is the identification of experiences to be included as cases in a distributed CBR system.

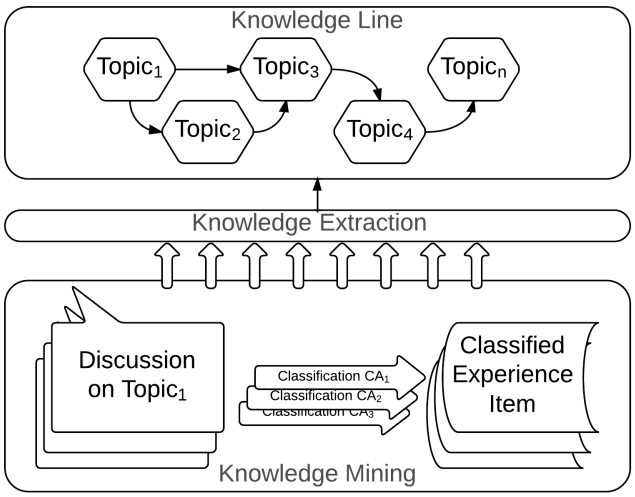

Fig. 1. Basic Knowledge Mining Approach in SEASALT

The Knowledge Mining component described in this paper is part of the SEASALT architecture [2]. SEASALT provides a general framework for creating a distributed knowledge-based system supporting the (semi-)automated identification, extraction and application of knowledge. Within SEASALT, we assume that unstructured text provided by users is available and should be populated into CBR systems. Therefore, we have created a methodology for the identification of distinctive topics that form a so called Knowledge Line [2]. A Knowledge Line describes a set of CBR-based agents, where each agent covers a topic and a solution is assembled by the partial solutions received from those agents (Topic Agents). Region, hospital, activity, person, disease, medicament, and potential risk are the topics of the travel medical application docQuery we will use as our running example in this paper. So, docQuery is a multi-agent system consisting of seven CBR systems as topic agents. The case base specific vocabulary is obtained from each agent's vocabulary knowledge container [16] and we will make use of the terms that have been modeled in the course of developing each agent. The work presented in this paper can be seen as a pre-processing step in which we are identifying relevant experience that are targeted to become cases eventually. Within the docQuery system, we have human Knowledge Engineers that build the cases as they are ensuring the quality of the cases. However, whithin SEASALT we are aiming at more and more supporting and automating this case

building task. Mining raw data and identifying relevant information is therefore an initial step.

Within the architecture we expecat to have one of more collector agents for each topic agent that monitor the user forum and trigger the knowledge extraction. When and how to trigger is the key task of the model-based classification presented in this paper. The remaining of this paper is structured as follows: In Section 2 we introduce the idea of Model-Based Knowledge Mining while Section 3 describes the supervised classification based on the knowledge models derived from CBR vocabulary as well as the SVM and C4.5 classifiers, which are a part of our agent ensemble. The following section compares the classification quality of these three agents in a real-life application in which forum discussions are classified. Section 5 discusses related approaches and the final section summarizes the findings and gives an outlook on future work.

2 Model-Based Knowledge Mining

The software agents, so called collector agents, require access to knowledge models that have been created during the implementation of the CBR systems as well as being a result of the Knowledge Extraction process [4]. Since we are mostly focusing on CBR systems, knowledge is represented as vocabulary (or gazetteers), cases, similarity measures, and optional transformation rules. The main knowledge containers we are using are the vocabulary for the gazetteer agent and the cases for learning the underlying models. Furthermore, we have included stop word lists for removing terms with less information.

CBR-Driven Vocabulary Within a SEASALT implementation, we create multiple, heterogeneous CBR systems, where each system has an individual case representation and vocabulary to cover the relevant cases. For example, the diseases case representation differs from an activity's case structure. We assume that the relevant vocabularies contain only those terms that are topic specific and characterize a particular domain. We will use this assumption to build software agents for each topic in order to extract relevant forum entries. In the remaining part of this section, we also assume that the CBR systems we created are using the myCBR tool. myCBR's SDK allows accessing the vocabulary per concept and attribute description [3]. We are able to receive all relevant terms, well organized and easy to distribute to the according agents. We decided to have one Gazetteer agent for each topic. The major task of the set of collector agents is identifying entries and organizing them in categories. Alongside the Gazetteer agent, we have also implemented C 4.5 and SVM agents, which use the keywords for learning the required models.

Stop Word Specific Vocabulary Before we can start the classification, we have to normalize the given texts, which in particular means removing stop words, based on stop word lists from the knowledge representation. We use both

German and English stop words since those are the languages we are currently dealing with, as well as HTML stop words. HTML stop words list is a manually created list of HTML tags occurring in the given data bases of forum entries. Also other frequently used terms in mailing lists should be removed in this preparation step. Stop word lists for the German and English language were retrieved from the *Wortschatzportal* of Leipzig University [14], from where they are available as plain text lists[3].

Knowledge Sources The instantiation we are currently focusing on, a web forum, is based on a mySQL server and therewith we can easily access the raw data inserted by forum users. The forum is restricted to experienced travelers, so we can assume they are experts in their domain. For that reason we will later on call this forum expert forum. Further on, we used the mySQL data base to store meta information, which has been automatically extracted, along with the manual and automatic classification for each forum entry. This enables us later to carry out various tests on the quality of the classification. The population of these parts will be described later on in this section. First we will introduce and characterize each type of agent.

Collector Agent Types For the collection and classification of forum posts we have three types of agents: Gazetteer agents, C4.5 agents and SVM agents. Since we aim to create modular and learning systems, we will furthermore have a supervisor agent that organizes each input of the basic classification agents and a third type, called learning or apprentice agent that monitors the actions of the Knowledge Engineer in order to provide feedback for the classifiers – or at least recognize if one of the collector agents fails permanently.

3 Supervised Classification in SEASALT

The SEASALT Knowledge Mining agents are realized based on the JADE framework [5] by first implementing the agent platform and then initializing the collector agents. The supervisor and Apprentice agents will be started a certain time after all collector agents are set up. The agent platform connects the software agents to the source data and the user interface, which can also start the agent platform.

For the startup of the supervisor agent, all collector agents are registered at the supervisor agent in order to receive data and monitor the actions carried out by the Knowledge Engineer on the forum entries or keyword lists.

Pre-processing of Forum Entries Before entries can be classified, they have to be normalized to reduce noise. Since we are dealing with natural language in the social web we decided for a case insensitivity approach and substitute all

[3] http://wortschatz.uni-leipzig.de/html/wliste.html

upper case letters by lower case letters as well as non-standardized characters are either removed or substituted. Finally multiple spaces are reduced to single spaces.

During the pre-processing of data to prepare the classification, we split longer texts into single sentences. From a longer forum discussion, we will receive a sentence as follows:

> [...] On the way to your hotel we already used the repellent to avoid mosquito bites. [...]

Later on each term will be handled as one element in an array, while each array contains a whole forum post by one user. Afterwards we carry out a first Named Entity detection for multi-word terms such as *Hepatitis A* or *Parkinson's Disease*, which should not be split up because this will cause a major loss of information. The example will then be represented as follows:

> [On] [the] [way] [to] [your] [hotel] [we] [already] [used] [the] [repellent] [to] [avoid] **[mosquito bites]**

Next, all stop words are removed and we have a resulting array containing potentially relevant terms. The example will then be represented as follows:

> [way] [hotel] [used] [repellent] [avoid] [mosquito bites]

Then we look up and tag each term with the topic class it belongs to. We then take for each keyword found n words before and behind and store them as our classification data. Later on, we will use this kind of term template to identify other, unknown terms describing the same or similar content. For $n = 3$ we will store the following data set

> way, hotel, used, <**keyword**>**repellent**< /**keyword**>, avoid, mosquito bites

with the association that this information entity serves the medication agent, which contains prevention information. Since we are working on a sentences base, we will not include terms from the next sentence.

Based in these entries we will train the intelligent classifiers. In Section 4, we are evaluating how many terms should be included to have an appropriate term template.

The overall goal is to collect experiences based on their description with which they are presented to others. For each topic or category, we are training the classifiers to recognize terms which are not included in our keyword list. This observation somehow creates a context in which keywords are used. This approach combines the boolean classification by the C4.5 and SVM agents with a probabilistic model, because we are trying, like Hidden Markov Model (HMM), to use surrounding information to derive classification for unknown terms. In comparison to HMM [9], which is based on probabilistic models, we use the C4.5 and SVM models. This approach can be compared to [8].

We perform this classification for each topic individually in order to receive independent classifications of the source data. This might lead to multiple classifications, which can be resolved by the Knowledge Engineer or confidence values. Currently we rely on the Knowledge Engineer in this regard. These steps are managed by the supervisor and Apprentice agent.

4 Experimental Evaluation

The evaluation of the knowledge mining has been carried out with two different data sets. The first one has been created manually from a Knowledge Engineer while the second model has been created semi-automatically. The creation of the semi-automatic model has been described in [4]. Each agent has been used in combination with the automatically and manually obtained knowledge models.

The goal of the evaluation is to find out which of the three implemented collector agents performs best in the given domain as well as how the two knowledge models work within our Knowledge Mining approach.

The data set has been created using the expert web forum with 700 entries in German. For training of the SVM and C4.5 agents we have used 200 entries and for the evaluation we took 500 test entries. From previous tests we learned by experience that taking into account a certain number of surrounding words, i.e., five words before and after a keyword, returns the best results [1], because 7 words were usually too many and sentence delimiters shortened the sequence, while 3 words did not produce stable results. Further on, we decided to use the standard SVM and C4.5 classifiers as they are available in WEKA.

For the evaluation we always used the complete Knowledge Models for the Gazetteer Agent and the SVM and C4.5 has been trained once. The variable factor is the noise in the incoming data in terms of stop words, which reduce the density of keywords (see Section 3). In the course of the evaluation we have used the three different kinds of stop word lists: stop word lists containing 100, 1,000 and 10,000 terms. In each run we collected the suggested classifications until 500 entries have been reached.

Figure 2 shows the F1 measures for the Knowledge Mining process using the Gazetteer agents for diseases, regions and medications. The complete results, which have been used to determine the F1 measure, can be found [2]. The F1 measures in this figure show the results for both models and there is a clear tendency that the Gazetteer agent performs much better than the SVM and C4.5 agent, respectively. As expected the model manually created by the Knowledge Engineer (first set of charts in Figure 2) is more reliable in the classification of new entries than the automatically created model.

Since the vocabulary of each CBR agent contains the terms relevant for representing the cases, the accuracy of the Gazetteer Agent is very high. The performance of the SVM and C4.5 agents turn out to be on the same level, while the SVM performs slightly better.

Overall, our experiments show that knowledge available in the vocabulary can be successfully used to classify unknown data during the pre-processing of

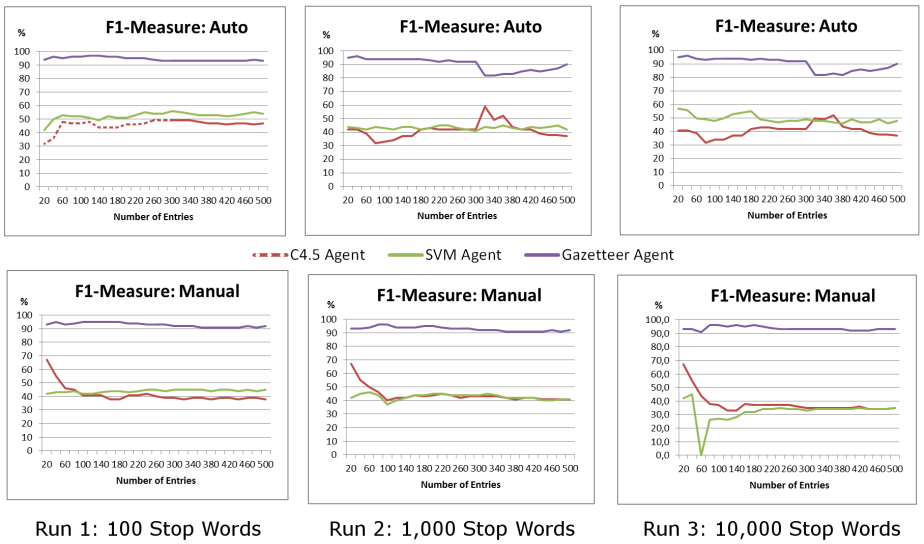

Fig. 2. F 1 Measures for the Diseases, Regions and Medication Agents

WWW resources in order to populate cases. However, this only enables a good classification, while the capturing of cases will be a different challenge.

5 Related Work

A related approach has been presented by Garcia and Wiratunga [15] in the context of Textual Case-Based Reasoning, where an unsupervised approach of learning taxonomies from web sources was introduced. However, our work can still be seen as a pre-processing step for the distributed CBR-driven multi-agent system, while their approach is directly applied within the CBR system without any human interaction. Similarly, Roth-Berghofer et.al. [17] used the vocabulary knowledge container to automatically index cases. We have taken this approach into account, and further developed these ideas away from the required rather static case structure to highly flexible and distributable case representations. Further, Zhang and Lesser [19] also address an hierarchical organisation of agents for distributed content sharing. However, their motivation is improving the performance of the computation, while our approach focuses on specialisation of tasks and content-based clustering of topics.

An alternative to the implemented knowledge mining approach could be making use of SMILA, an architecture specialized for the search in unstructured information sources. SMILA has been developed as middleware platform within the Theseus program - mainly for the application scenario ORDO[4]. SMILA is based on the OSGi framework [18,11].

[4] `http://www.theseus-programm.de/anwendungsszenarien/ordo/default.aspx`

The architecture is divided in two parts: pre-processing and the search engine itself. Since SMILA heavily uses OSGi's service components it contains various individually configurable modules [10]. The *Pre-Processing* uses agents or services for crawling and processing unstructured information in order to build an index that can be searched afterwards. The main contribution of SMILA is the provision of an open middleware that has to be further developed.

The development in SEASALT and SMILA was carried out in parallel and at an early stage the middleware did not meet our expectations regarding a very strict indexing and searching focus rather than a high variability of information and knowledge processing. Today, after SMILA is an active project within the eclipse foundation an integration of our modules is more feasible and SEASALT could benefit from SMILA's performance when dealing with big data. Since the main focus is searching the provided processes are tailored in this way, however as shown in [7], SMILA can also be used in various ways such as for dealing with more structured sources and carrying out more sophisticated tasks like providing adaptation capabilities.

Further on, rather than including knowledge models from myCBR, also Protégé [13,6] would be an option if just ontologies are to be included. We have worked with both, but eventually decided for myCBR since we are focusing on CBR-driven applications. Ontologies modeled with the *Protégé-Frames Editor* are also accessible from our tool [1].

6 Conclusion and Outlook

The work presented in this paper targets at reusing the vocabulary knowledge container for classifying new entries whether they fit in the topic of existing CBR systems. The approach has been implemented as SEASALT instance [2]. SEASALT as well as the introduced Knowledge Mining approach have been applied in the real-life application docQuery and the data used for the evaluation of our work was obtained from an expert forum in travel medicine. The experiments show that the pre-processing and selection of web-data can be based on the knowledge created in CBR systems as the gazetteer agents, which are based on various CBR system's vocabularies, outperform standard Machine Learning approaches. Moreover, the effort of creating the Gazetteer agents is very low, since they directly use the knowledge models provided by myCBR. In contrast, training data for the SVM and C4.5 classifiers has to be created before the classifier can be applied.

The Knowledge Mining approach presented in this paper offers a new, pragmatic perspective for constructing WebCBR systems [12] with a positive cost-benefit relationship. Moreover, the compatibility to SMILA can be used to create more parallel knowledge mining approaches which will enable a more effective creation of CBR systems capturing cases from web resources. Also, up to now, we only use the plain keywords rather than the complete taxonomies for the classification. A direction we will investigate further is the development of case-based

classifiers, which can be directly derived from each CBR agent in the Knowledge Line.

Acknowledgement We would like to thank our students Kirsten Skibbe, Manuel Ahlgrim, and Alena Rudz for their contributions to the work presented in this paper.

References

1. Ahlgrim, M.: Developing software agents for experience classification from web forums (Analyse von Webcommunities und Extraktion von Wissen aus Communitydaten für Case-Based Reasoning Systeme). University of Hildesheim, Master thesis (December 2010)
2. Bach, K.: Knowledge Acquisition for Case-Based Reasoning Systems. Ph.D. thesis, University of Hildesheim (2013)
3. Bach, K., Althoff, K.D.: Developing Case-Based Reasoning Applications Using myCBR 3. In: Watson, I., Agudo, B.D. (eds.) Case-based Reasoning in Research and Development, Proceedings of the 20th International Conference on Case-Based Reasoning (ICCBR-12). pp. 17–31. LNAI 6880, Springer (September 2012)
4. Bach, K., Sauer, C.S., Althoff, K.D.: Deriving case base vocabulary from web community data. In: Marling, C. (ed.) ICCBR-2010 Workshop Proc.: Workshop on Reasonng From Experiences On The Web. pp. 111–120 (2010)
5. Bellifemine, F., Caire, G., Trucco, T., Rimassa, G.: JADE Programmer's Guide. CSELT S.p.A., TILab S.p.A., Telecom Italia S.p.A., o.O. (2010)
6. Gennari, J.H., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubzy, M., Eriksson, H., Noy, N.F., Tu, S.W.: The evolution of protg: An environment for knowledge-based systems development. International Journal of Human-Computer Studies 58, 89 – 123 (2002), http://bmir.stanford.edu/file_asset/index.php/52/BMIR-2002-0943.pdf
7. Hanft, A., Schäfer, O., Althoff, K.D.: Integration of drools into an osgi-based bpm-platform for cbr. In: Agudo, B.D., Cordier, A. (eds.) ICCBR-2011 Workshop Proceedings: Process-Oriented CBR (2011)
8. Joachims, T., Finley, T., Yu, C.N.: Cutting-plane training of structural svms. Machine Learning 77(1), 27–59 (2009)
9. Marsland, S.: Machine Learning - An Algorithmic Perspective. Chapman and Hall/CRC Machin Learning and Pattern Recognition Series, CRC Press, Taylor and Francis Group, Boca Raton, Florida, USA (2009)
10. Nieland, D., Stiglic, J.: Open services gateway initiative (osgi) to drive development of gateway standard for homes, soho and remote locations. Palo Alto: Open Services Gateway initiative, Palo Alto, Californien, USA (November 1999), http://www.osgi.org/News/19991122EN
11. Novakovic, I.: Smila/architecture overview. Ottawa: Eclipse Foundation, Inc., Ottawa, Ontario, Canada and Empolis GmbH (Januar 2010), http://wiki.eclipse.org/SMILA
12. Plaza, E.: Semantics and experience in the future web. In: ECCBR '08: Proceedings of the 9th European conference on Advances in Case-Based Reasoning. pp. 44–58. Springer-Verlag, Berlin, Heidelberg (2008)
13. Protégé: what is protégé? Stanford : Stanford University School of Medicine, Stanford Center for Biomedical Informatics Research (2012), http://protege.stanford.edu/overview/index.html

14. Quasthoff, U., Richter, M.: Projekt deutscher wortschatz. Babylonia (2005)
15. Recio-Garcia, J.A., Wiratunga, N.: Taxonomic semantic indexing for textual case-based reasoning. In: Proceedings of the 18th international conference on Case-Based Reasoning Research and Development. pp. 302–316. ICCBR'10, Springer-Verlag, Berlin, Heidelberg (2010)
16. 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)
17. Roth-Berghofer, T., Adrian, B., Dengel, A.: Case acquisition from text: Ontology-based information extraction with scoobie for mycbr. In: Bichindaritz, I., Montani, S. (eds.) Case-Based Reasoning. Research and Development, Lecture Notes in Computer Science, vol. 6176, pp. 451–464. Springer Berlin Heidelberg (2010)
18. Schütz, T.: D11.1.1.b: Concept and design of the integration framework. Bundesministerium für Wirtschaft und Technologie and Theseus-Ordo and Empolis GmbH (2008)
19. Zhang, H., Lesser, V.: A Dynamically Formed Hierarchical Agent Organization for a Distributed Content Sharing System . In: Proceedings of the International Conference on Intelligent Agent Technology (IAT 2004). pp. 169–175. IEEE Computer Society, Beijing (2004), `http://mas.cs.umass.edu/paper/373`