# Multi Agent Knowledge Acquisition for the SEASALT Apprentice Agent using Twitter Feeds

Chathuri Nilushika Seneviratne, Christian Severin Sauer, and Thomas Roth-Berghofer

School of Computing and Technology, University of West London,
St Mary's Road, London W5 5RF, United Kingdom
{first.lastname}@uwl.ac.uk

**Abstract.** Web 2.0 content, including blogs, forum posts and tweets, is mostly expressed in an unsystematic manner. Due to this reason, retrieving and reusing this content has become challenging. As a solution, Reichle et al. [6] presented a novel architecture named SEASALT. A main feature of the SEASALT architecture is the use of topic agents implemented as topic specific Case-based reasoning (CBR) systems. Another key component of the SEASALT architecture is the Apprentice Agent which supports a knowledge engineer in the SEASALT Architecture by automatically extracting vocabulary items and taxonomic similarity measures for the CBR-based topic agents from a virtual community of experts, comprising the knowledge input for the SEASALT Architecture. A first implementation of such an apprentice agent was presented by Bach et al. [6] with the Knowledge Extraction workbench (KEWo) which extracted vocabulary items and similarity measures form an online community of travel medicine experts. The work presented in this paper extent the KEWo to use Twitter feeds as a knowledge source. A Multi Agent System is developed to acquire Twitter feeds which are then transferred for further knowledge extraction to the Apprentice agent component KEWo. Further Twitter is analysed as a knowledge source in terms of the amount of data it can provide on a specific topic and how this provided amount of tweets has an impact on the performance and quality of knowledge extracted from them. Furthermore, the paper analyses how well the hash tag feature provided in Twitter can be employed as a source of structuring information. As the ultimate output, this paper contributes to the extension of a *virtual community* within the SEASALT architecture by including the content from Twitter users.

**Keywords:** CBR, Knowledge Extraction, Twitter, Similarity Measures, SEASALTexp

## 1 Introduction

With the Web 2.0, the static and well-structured web content is progressively replaced by individually sometimes very loosely structured user-generated content

such as blogs, forum posts or tweets. This user-generated content often contains user experiences which are mostly expressed in an unsystematic manner. Hence, retrieving and reusing this content in an automated way is challenging. Furthermore, the traditional approaches like monolithic databases or text mining techniques are not sufficient to deal with the wealth of experiences provided in today's World Wide Web [6]. As a solution, Reichle et al. [6] presented SEASALT, "a novel architecture for extracting, analysing, sharing and providing community experiences" drawing on these volatile web sources.

The domain of travel medicine is an interdisciplinary speciality concerned with health problems associated with travel that covers all medical aspects that a traveller has to deal with before, during and after a journey [8]. As a test bed instantiation of SEASALT architecture the docQuery [6] project was developed to provide recommendations and advise to its users within the travel medicine domain. This paper describes an expansion to the existing docQuery project that enables the docQuery project to use Twitter-feeds or tweets as a knowledge source. In this paper we detail on how we extended the docQuery's Apprentice Agent component KEWo [1] to enable it to acquired Twitter feeds related to travel medicine employing a multi-agent system that automatically extracts vocabulary items and similarity measures form these tweets. The extracted vocabulary items and taxonomic similarity measures can then be used by the CBR systems employed by the Topic Agents of the docQuery system.

The rest of this paper is structure as follows: We introduce the research goals that we answered within our work presented in this paper in section 2. We then detail on the work related to our research in section 3. After analysing the technologies involved in our research in section 4 we move on to detail on the implementation of our prototype software in section 5. We then detail on the evaluation and performance of our prototype software in section 6. A summary and conclusion7 then conclude the paper.

## 2 Aims and Opportunities of our work

This aim of the research work described in this paper was to answer the following four research questions:

1. Can Twitter provide a sufficient amount of data on a specific topic, in a specific domain to serve as raw material for information/knowledge extraction within the KEWo?
2. How well do hash tags within Tweets perform as a source of structuring information within a collection of tweets?
3. How does the volume of tweets analysed impacts on the performance and quality of the knowledge extraction?
4. Can some form of provenance information/quality measure be extracted/ applied out of/to a collection of Tweets serving as a raw text for Knowledge extraction within the KEWo?

The opportunities intend to gain from answering the above research questions are manifold. Being able to access twitter feeds as sources of knowledge for our

systems enables the system to reason on almost real-time knowledge as well as to acquire access to a vast amount of this knowledge. Being able to establish quality measures for the tweets within the twitter feeds would be beneficial with regard to the selection of good quality raw knowledge that is worthwhile the effort of acquiring and formalising it as we assumed that meta information, such as a Twitter users number of followers and the number of re-tweets of a tweet could be used as provenance information of the raw knowledge to guide the automatic decision of whether to include a tweet in the extraction process or not. Gaining insight on the presence and use of structuring knowledge present in tweets would benefit the re-use of this knowledge in the knowledge formalisation process realised within the KEWo, thus reducing the computational effort of said process. Furthermore, besides the specific use of the extended KEWo un the docQuery context, the implementation of a generic extraction approach of vocabulary items and taxonomic similarity measures for these, would enable CBR systems in general to acquire knowledge from Twitter feeds.

## 3 Related work

docQuery [6] is a medical information system for travellers based on the SEASALT architecture. The purpose of docQuery is to provide advise on travel medicine. Once the user enters key data of the travel such as destination, travel period and age of travellers, into the system, the system will then produce a leaflet advising the traveller on relevant medical information and actions, such as required vaccinations, for his/her specific trip. To generate these leaflets docQuery implements the Knowledge line approaches by splitting up information retrieval to a number of Topic Agents, each being a CBR system handling a specific topic of the query, such as specific locations, disease or medications. To provide the knowledge on which the Topic agnt's CBR systems operate docQuery initially mined a web forum consisting of travel medicine experts. The Mining or knowledge gathering and formalisation task was carried out by a knowledge engineer with the aid of an apprentice agent component KEWo, see section 4 for a more detailed description of the KEWo.

As mentioned, following the SEASALT architecture, the docQuery project employs eight different CBR systems, each of which represent a certain topic agent. The work carried out by Sauer et al. [8] focuses on one of these CBR systems which contain information about diseases related to travel medicine. They proposed an approach to extract knowledge from Linked Open Data (LOD) sources and to integrate in a CBR system. In this approach data extracted from LOD is further refined using KEWo to generate taxonomies.

A further development of the SEASALT architecture can be found in the work of Bach et al. [1], which is focused on extracting knowledge for CBR systems from web based communities. In their work, they present a process for knowledge extraction from web communities which can be applied to extract knowledge for CBR systems. Figure 1 shows the proposed knowledge extraction process. In addition, the paper further discusses applying the introduced knowl-

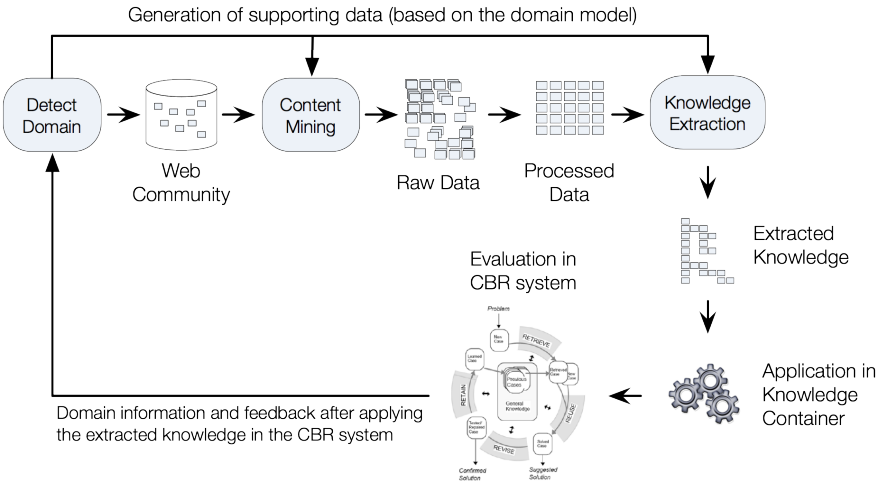


**Fig. 1.** Process for knowledge extraction from web communities[1]

edge extraction process in a real life application for the travel medicine domain based on web forum data. They have utilized KEWo to extract taxonomic similarity measures from web forum data as well as to derive vocabulary items. According to Bach et al. [1] KEWo supports the gathering, pre-processing and extraction steps of the process.

The related work discussed so far denotes knowledge acquisition from web based sources such as forum posts and LOD. Our research focussed on Twitter feeds as a knowledge source within the SEASALT architecture, expanding the virtual community of expert by the input from Twitter feeds, as depicted in Figure 2 2). There are already a number of studies, analysing Twitter feeds as a knowledge source. An empirical study of topic modelling in Twitter is conducted by Hong et al. [4] as opposed to using standard text mining tools to analyse micro-blogging content. In addition to that Wang et al. [10] investigated a sentiment classification in Twitter, based on hash tags. A further study carried out by Cheong et al. [2] investigate on web-based intelligence retrieval and decision making from the Twitter trends knowledge base as opposed to traditional blog analysis. The research of Varga et al. [9], detects topics of tweets using DBpedia and Freebase. Mendoza et al. [5] analyse in their study, how information is propagated through Twitter network with the purpose of assessing its reliability as an information source. They state that rumours are questioned more than news in the Twitter community, which makes it possible to detect rumours by using aggregate analysis on tweets [5].

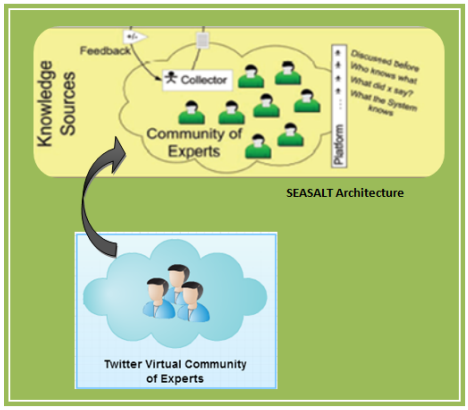


**Fig. 2.** Twitter Virtual Community of Experts within SEASALT Architecture

## 4 Technologies involved in our reserach

In this section we examine the three major technologies involved in our research. These technologies were the SEASALT Architecture and it's current implementation of an apprentice agent, namely the KEWo and the Twitter platform technology and its API as being the technology that is used to acquire the knowledge from Twitter feeds.

### 4.1 SEASALT Architecture

SEASALT (Sharing Experience using an Agent-based System Architecture Layout) [Figure 3] is an application-independent architecture featuring knowledge acquisition from web communities, knowledge modularization and agent-based knowledge maintenance [6]. The SEASALT architecture employs a knowledge line approach which represents a modularization of knowledge by breaking down a complex topic into sub topics handled by topic agents. In the SEASALT architecture topic agents can be any kind of information system or service such as CBR systems, data bases or web services [6]. This architecture consists of several components where a detailed explanation is provided in the work of Reichle et al. [6]. In the docQuery project the topic agents are implemented as eight different CBR systems for which the KEWo, as an implementation of an apprentice agent, is extraction vocabulary items as well as taxonomic similarity measures. The KEWo is extracting this knowledge from the community of experts, consisting of a web forum of travel medicine experts.

### 4.2 Knowledge Extraction Workbench (KEWo)

The Knowledge Extraction Workbench (KEWo) is an implementation of the apprentice agent described within the SEASALT Architecture. The KEWo (ap-

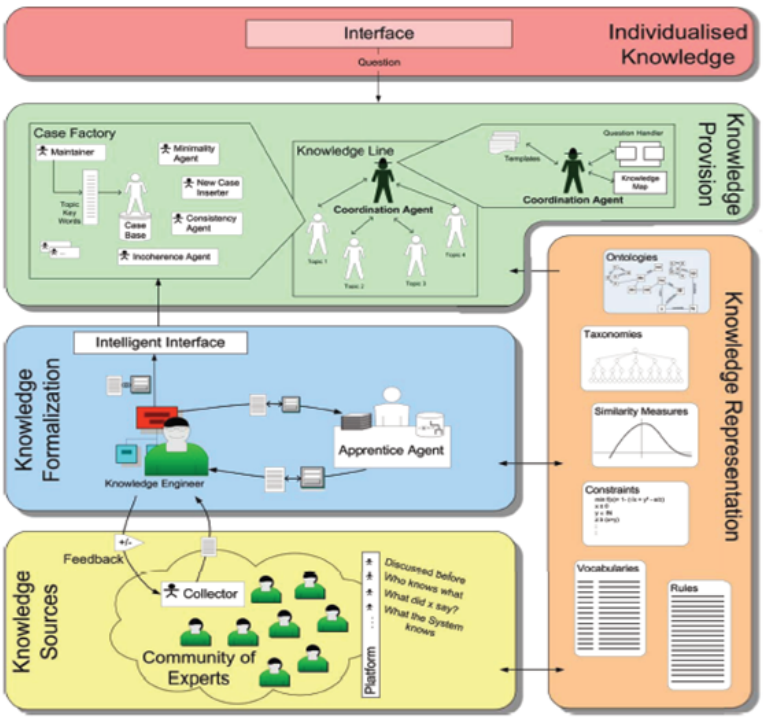


**Fig. 3.** The SEASALT Architecture [6]

prentice agent) is used in the knowledge formalization phase of the SEASALT architecture.

The KEWo itself is a Java-based middleware for knowledge extraction for CBR systems. KEWo extracts vocabulary items for a specific domain and generates taxonomies of these by the calculation of distance-based similarity measures between the vocabulary items within the taxonomies. The underlying extraction process for symbols is implemented by the use of the information extraction tool ANNIE (A Nearly-New Information Extraction System), being a part of the GATE framework for natural language processing. The KEWo employs a customised for ANNIE, using specific gazetteers and rule sets that identify terms from a specific domain. Thus KEWo is able to extract symbols from unstructured texts and either build completely new or expand existing taxonomies of symbols to be used in docQuery's topic agent's CBR systems. The KEWo as an implementation of an apprentice agent is currently extracting Vocabulary items (terms of diseases, locations and medicaments) as well as similarity measures in the form of taxonomies of these terms, generated from web community data. Initial development of KEWo was carried out by Sauer [7] to analyse natural language forum postings in a web community of travel medicine experts.

### 4.3 Twitter

Twitter is an online social networking and micro-blogging service that allows its users to post and read messages, known as tweets. In 2013, Twitter reported to be celebrating its seventh birthday with 200 million users worldwide who send an average of 400 million tweets everyday [3]. Due to its instantaneous nature and ease of use [11] currently it has taken the form of a conversational blogging and an online social network [2]. Varga et al. [9] states that micro-blogging platforms such as Twitter serve as a real-time information channel, which contains rapidly up-to-date information on verity of topics compared to other traditional news sources.


**Fig. 4.** Example tweets relating to the domain of travel medicine

As tweets are limited to 140 characters, users invented many techniques to expand the semantics that are carried out by these messages [4].For instance, URL shortening services (e.g., http://www.bit.ly) are often used by Twitter users when posting external URLs in their tweets. In addition, frequent use of abbreviation can also be seen as a result of the limited size of a tweet. These characteristics are to be taken into account while aiming to extract knowledge from tweets. On the other hand, tweets often contain hash tags (starting with # sign), which are frequently used to identify events or topics created by users. These hash tags are highly user driven and often contain multiple words without separating blanks (e.g. denguefever) or using underscore sign to separate words (e.g., dengue_fever). Another challenge for the text mining from tweets is the use of non-standard English and frequent misspellings and use of jargon [9]. However, Mendoza et al. [4] state that tweets may convey very rich meanings, even though the content of messages is limited.

## 5 Prototype implementation

In this section we are demonstrating the initial implementation of our prototype. We detail on the use of the Twitter API to access the Twitter feeds as well as on

our implementation of the gathering agents and the knowledge representation we chose for our prototype implementation.

### 5.1 Implementation

Our application is designed to be independent from both Twitter and KEWo, which enables this application to use other knowledge sources to retrieve information and other feeders to further analyse the source information. Our application consists of four modules, namely; Domain, Feeder, Integration and Source. The Domain module contains POJOs (Plain Old Java Object) used in the application. One domain class is used to set the query condition, such as the search text, search language etc. A second domain class is used to set the source data such as search text and the data feed. The Feeder Module consists of the Feeder Connector which provides the basic interface for knowledge extraction applications such as KEWo to get connected to the source data. Any number of feeders can be plugged in to this architecture by implementing the Feeder Connector interface. The Source module consists of the Source Connector which provides the basic interface to get connected with knowledge sources such as Twitter, using the Twitter API. Any number of source connectors can be plugged in to this architecture by implementing the Source Connector interface. Finally the Integration Module is used to connect both the source and feeder. In this implementation the integration module connects the Twitter feeds to KEWo. The integration module also consists of the agents used to represent the domains diseases, locations and medicaments. However, these agents are not displayed in figure 5 as it would make it to complex. So in our application Agents connect with the Source Connector module in order to perform searches on Twitter feeds and then connected to Feeder Connector modules to transfer the tweets into the KEWo for further knowledge extraction.

### 5.2 Twitter *Group-By Features*

In order to improve the analysis of Tweets, certain *Group-By Features* are implemented in our prototype. The *Group-By Features* group together tweets on certain criteria before it is transferred to KEWo for further analysis. The first criteria implemented for this *Group-By Feature* is the number of re-tweets. Re-tweet is the option that allows a Twitter user to repost a tweet that has been posted by another user. Number of re-tweets was chosen as a *Group-By Feature* based on the assumption that a re-tweet can possibly happen when the message content has some value in it. For example, if it contains mundane, such as what I had for breakfast it is highly unlikely that a user will re-tweet such content unless the person who tweeted that message was a celebrity. The second criteria used as a *Group-By Feature* is the number of followers the author of a tweet has. A user has greater number of followers, denotes that there are lot of people following that person in Twitter. Hence, it is assumed that such person will post tweets with some valuable content, as there are many people following that user. After arranging the tweets with the *Group-By Features*, all the Twitter feeds are


**Fig. 5.** The prototype application structure

transferred to KEWo for further analysis. For a certain keyword, it is possible to analyse all the available tweets through KEWo, as well as the tweets grouped by according to number of followers and number of times a post has been re-tweeted. By introducing these *Group-By Features*, it was expected to create a virtual community of experts using Twitter to cater SEASALT architecture as a knowledge source (Figure 2).

### 5.3 Agent Implementation

The Multi Agent System developed in this work is based on the JADE framework. Basically, it provides two methods of initiating agents; manually at compile time and dynamically at run time. In our prototype dynamic run time instantiation of JADE agents are used. Three agents are implemented, one each for the domains of disease, location and medicament.

### 5.4 Knowledge representation within the Prototype Implementation

The representation of the raw knowledge gathered from the Twitter feeds was realised using a MySQL database. The database's structure was based on the

initial database structure employed by the KEWo to enable the KEWo to 'read' and then process the gathered raw data with minimal adaption effort.

## 6 Experiments and Evaluation

Our initial simple approach to identifying tweets in the domains of disease, location and medicament was to specify 10 keywords for each domain that controlled our twitter feed gathering for each domain. We then analysed whether Twitter provides sufficient amount of data on a specific topic, in a specific domain. This analysis resulted in identifying that within the gathered tweets of one week, location related tweets provided the highest amount of data with an average of 2200 tweets, followed by disease related tweets with an average of 1250 tweets and medicament with an average of 440 tweets per keyword. As an example, retrieving tweets over the time of one week that mentioned the keyword *Rabies* returned 2599 tweets.

Due to the lack of data available in the medicaments domain, it was recommended to gather tweets by gathering Twitter feeds over the time of several weeks to accumulate sufficient raw knowledge on medicaments. The gathered tweets were then analysed using the KEWo, which generated taxonomies with sufficient amounts of detail, see figure 6 for an example of a taxonomy built upon the search term 'Aspirin'. Providing a greater amount of 'raw' tweets as material for the KEWo to extract from the taxonomies generated were deeper and had a rich amount of child nodes whereas short and flat structured taxonomies were created when a lower amount of tweets were transferred to the KEWo for information extraction.

It is important to consider how to address the limitation of Twitter API that limits the search to the timespan of one week. As a solution this project maintains a data store that allows the application to save the returned tweet objects for every search performed. Through this solution, it is expected to gather a vast amount of data on each domain, by running the search for a couple of weeks. An interesting finding of our research was that hash tags perform very well as a source of structuring information within a collection of tweets. The evaluation of our experiment's results indicates that locations are more often hash tagged, compared to the tweets in other domains. Abbreviated terms such as **HIV** and **TB** were often hash tagged in tweets. Apart from this fact, terms related to diseases and medicaments were showing a lesser probability of being hash tagged in tweets. The experiments carried out also indicate that within the location and diseases domains hash tags served as a good source of structuring information.

With regard to the extracted taxonomies, although the volume of tweets had an impact on the taxonomy structure and length, some taxonomies contained irrelevant terms regardless the volume of analysed tweets. Hence, we introduced a set of *Group-By Features* in order to create a Virtual Community of Experts within the SEASALT architecture, using the quality and provenance information from the raw tweets to allow the KEWo to focus on tweets that were estimated, based on the provenance information, to be from experts in the domain of

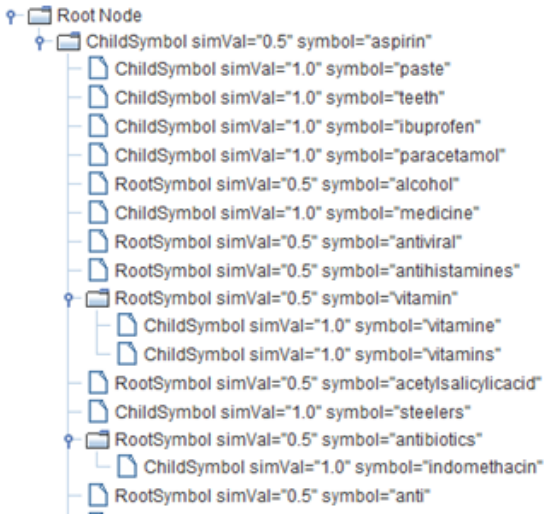


**Fig. 6.** Example taxonomy extracted for the search term 'Aspirin'

travel medicine. It was observed that the tweets having a re-tweet count ranging from 0-10 generated taxonomies of better quality compared to other tweets. We measured the quality of the taxonomies generated by manually inspecting the parent-child relationship of vocabulary items in the taxonomies with regard to being correct relationships. It was also found that the *Group-By Feature* based on the number of followers did not facilitate identifying high quality raw knowledge tweets. However just using this *Group-By Feature* still enabled the KEWo to generate taxonomies with a high amount of correct parent-child relationships of the vocabulary items in the taxonomies.

## 7 Summary and Outlook

This paper presented a "'Virtual Community of Experts"' using Twitter as the knowledge source within the SEASALT architecture as well as two *Group-By Features* to establish high quality tweets as raw knowledge. A multi agent system that represents the domains of diseases, locations and medicaments was developed to acquire Twitter feeds in the respective domains. The gathered raw knowledge from the tweets was then fed to the KEWo, an instantiation of SEASALT Apprentice Agent. We were able to prove Twitter as an applicable knowledge source within the SEASALT architecture and that it is possible to extract vocabulary items and taxonomic similarity measures, to be used within CBR systems, from Twitter feeds. We were able to do so by showing a sufficiently accurate knowledge extraction within the domains of diseases, locations and medicaments. Moreover we were able to establish that hash tags in tweets are a good source of structuring information within a collection of tweets, especially

in the domains of diseases and locations. Even though precise quality measures could not be derived from the introduced *Group-By Features*, the effort taken to create a Virtual Community of Experts for the SEASALT architecture, employing the provenance and quality information from gathered from our introduced *Group-By Features*, can be seen as the foundation for further enhancements in using Twitter as a knowledge source. For our immediate future work we plan to perform more experiments on broader domains than travel medicine and also aim for establishing more *Group-By Features* as quality measures to judge tweets before they are processed by the KEWo.

## References


1. Bach, K., Sauer, C.S., Althoff, K.D.: Deriving case base vocabulary from web community data pp. 111–120 (7 2010)
2. Cheong, M., Lee, V.: Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In: Proceedings of the 2nd ACM workshop on Social web search and mining. pp. 1–8. SWSM '09, ACM, New York, NY, USA (2009)
3. Elvin, L.: It's many happy retweets as Twitter reaches a milestone: Phenomenon: Networking website with 200 million users is seven years old. `http://search.proquest.com.ezproxy.uwl.ac.uk/docview/1322485410` (2013), accessed: 2013-08-10
4. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics. pp. 80–88. SOMA '10, ACM, New York, NY, USA (2010)
5. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: can we trust what we rt? In: Proceedings of the First Workshop on Social Media Analytics. pp. 71–79. SOMA '10, ACM, New York, NY, USA (2010)
6. Reichle, M., Bach, K., Althoff, K.D.: The seasalt architecture and its realization within the docquery project. In: Proceedings of the 32nd annual German conference on Advances in artificial intelligence. pp. 556–563. KI'09, Springer-Verlag, Berlin, Heidelberg (2009)
7. Sauer, C.: Analyse von Webcommunities und Extraktion von Wissen aus Communitydaten fr Case-Based Reasoning Systeme. Master's thesis, University of Hildesheim (2010)
8. Sauer, C.S., Bach, K., Althoff, K.D.: Integration of Linked Open Data in Case-Based Reasoning Systems. In: Atzmüller, M., Benz, D., Hotho, A., Stumme, G. (eds.) Proceedings of LWA2010 - Workshop-Woche: Lernen, Wissen & Adaptivitaet. Kassel, Germany (2010)
9. Varga, A., Cano, A.E., Ciravegna, F.: Exploring the similarity between social knowledge sources and twitter for cross-domain topic classification (2012)
10. Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 1031–1040. CIKM '11, ACM, New York, NY, USA (2011)
11. Williams, S.A., Terras, M.M., Warwick, C.: What do people study when they study twitter? classifying twitter related academic papers. Journal of Documentation 69(3), 384–410 (2013)