Knowledge acquisition for the SEASALT apprentice agent using Twitter feeds

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. The recent developments of Web 2.0, has driven the web content from its static and formalised nature to a highly user-driven nature. Such web content includes blogs, forum posts and tweets which are 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] present a novel architecture named SEASALT and within this architecture present the docQuery project, carried out as one instantiation of the presented architecture focusing on the domain of travel medicine. The work presented in this paper is demonstrating the use of Twitter feeds as a knowledge source within the SEASALT architecture, expanding the knowledge-base of the docQuery project. A Multi Agent System is developed to acquire Twitter feeds related to travel medicine, which are then transferred for further knowledge extraction to the Apprentice agent component of the SEASALT Architecture named: Knowledge Extraction Workbench (KEWo). In this paper, 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 a result of this analysis, a set of *Group-By Features* is introduced to enhance the knowledge extraction based on attributes of Twitter feeds such as *retweet count* and *number of followers*. As its final output, this paper demonstrates how to create a *virtual community of experts* within the SEASALT architecture for further knowledge extraction from said community.

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

1 Introduction

With the development and expansion of 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 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] present 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 acquired Twitter feeds related to travel medicine employing a multi-agent system based processing a user query. Further we demonstrate how the acquired tweets are then processed by the Knowledge Extraction Workbench (KEWo) [1] which extracts vocabulary items as well as similarity measures from the acquired tweets.

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 concludes 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. 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.

3 Related work

docQuery [6] is a medical information system for travellers based on the SEASALT architecture. Once the user enters key data of the travel such as destination, travel period and age of travellers, the system will produce a leaflet containing individually composed information that the traveller has to be aware of in terms of travel medicine for the requested query. A forum consisting of professionals and experts of travel medicine domain is used as the knowledge source for docQuery project. Furthermore, knowledge formalisation is carried out by a knowledge engineer with the aid of an apprentice agent developed using GATE and RapidMiner[1] [6]

According to SEASALT architecture docQuery project consists of 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 propose 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 knowledge extraction process which can be applied for all kinds of knowledge based systems including CBR systems. The proposed knowledge extraction process is as follows;

1. Domain detection
2. Web community selection
3. Content mining
4. Processing raw data
5. Processed data
6. Knowledge extraction
7. Extracted knowledge
8. Application in knowledge container
9. Evaluation

In addition, the paper further discusses applying the introduced knowledge extraction process in a real life application for the travel medicine domain based

[1] RapidMiner: open source system for data mining.

on web forum data. They have utilized KEWo to extract taxonomies from web forum data to derive case base vocabulary. According to Bach et al. [1] KEWo supports the steps 4 to 7 of the knowledge extraction process.

The related work discussed so far denotes knowledge acquisition from web based sources such as forum posts and LOD. Even though Twitter feeds have been noted as possible web based knowledge source consisting user-driven information, no initiative had yet been taken to instantiate knowledge acquisition from tweets within the SEASALT architecture. Hence, our research was initiated to use Twitter feeds as a knowledge source within the SEASALT architecture (Figure 1).

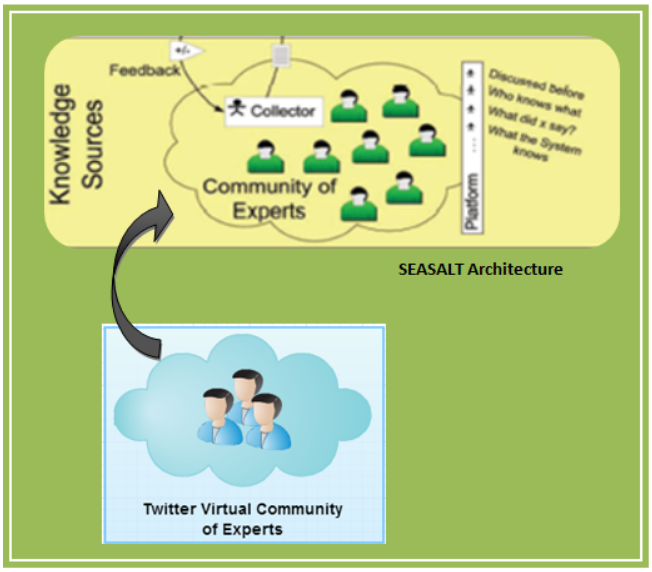

Fig. 1. Twitter Virtual Community of Experts within SEASALT Architecture

However, there are number of studies performed on 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] propose a sentiment classification in Twitter based on hash tags. The sentiment polarity of tweets containing the hash tag, hash tag co-occurrence relationship and the literal meaning of hash tags are the three types of information that has been used in this investigation to analyse Twitter feeds. Furthermore, a 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. Also they investigate the similarity between these knowledge sources and Twitter at the lexical and conceptual level. Moreover, 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].

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 2] 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] For our research we extended an existing prototype software, docQuery wihich is an instance of the SEASALT Architecture.

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 (apprentice agent) is used in the knowledge formalization phase of the SEASALT architecture. The apprentice agent aids the knowledge engineer by creating and presenting the results of the (automatic) knowledge extraction to her. The knowledge engineer can then deem the extracted knowledge as acceptable or amend or reject it. 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.

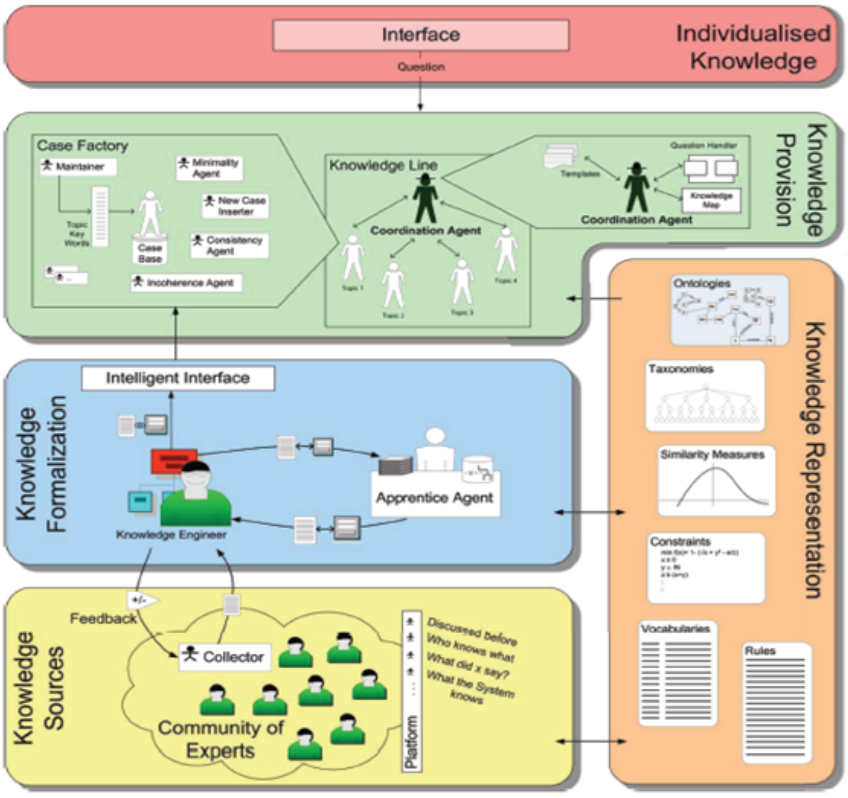

Fig. 2. The SEASALT Architecture [6]

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]. Initially, Twitter was started with the intention of sharing the answer to the question of what am I doing, with family, friends and acquaintance. However, 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] for sharing news or reports about events, ranging from mundane through emerging information about social political situations to emergencies [11]. 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.

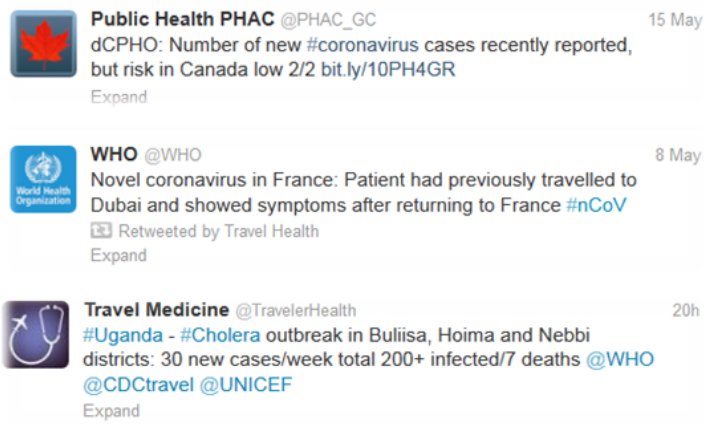

Fig. 3. Example tweets relating to the domain of travel medicine

There are several challenges of using standard text mining tools to acquire knowledge from tweets. One of the challenges is the restricted size of a tweet. 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. On the other hand, tweets often contain hash tags (starting with # sign), which are frequently used to identify events or topics created by users. However, these hash tags are highly user-driven and it may 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 with 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 Twitter API

An Application Programming Interface (API) provides a set of programming instructions and standards for a program to accomplish a task. Software companies such as Twitter make their APIs available to public so that application

developers can design products empowered by their services. Programmers use Twitter API to develop websites, widgets, applications and other Twitter related projects. Twitter API is entirely HTTP-based and it consists of two segments; the REST API and Streaming API.

Fig. 4. The prototype application structure

5.2 Twitter *Group-By Features*

In order to improve the analysis of Tweets, certain *Group-By Features* are implemented in our prototype. As it is explained by the name, these *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 retweets. Retweet is the option that allows a Twitter user to repost a tweet that has been posted by another user. Number of retweets was chosen as a *Group-By Feature* based on the assumption that a retweet 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 retweet 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 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 retweeted. 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 1).

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. Figure 5 provides an overview of the structure of the database we implemented to store the raw knowledge gathered from the Twitter feeds. The database structure is developed 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

Initially, we 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 whereas disease related tweets represented a considerable amount of data. However, 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.

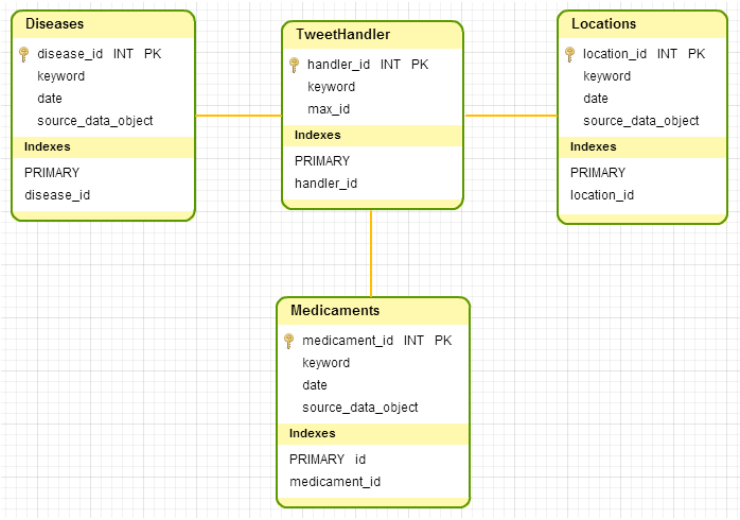

Fig. 5. Data base representation of the gathered raw knowledge

Another major finding of our research was how well the hash tags perform as a source of structuring information within a collection of tweets. The evaluation of our experiment's results indicate that locations are more often hash tagged, compared to the tweets in other domains. Also 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 travel medicine. It was observed that the tweets having a retweet count ranging from 0-10 generated taxonomies of better quality compared to other tweets. 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 reasonably accurate taxonomies.

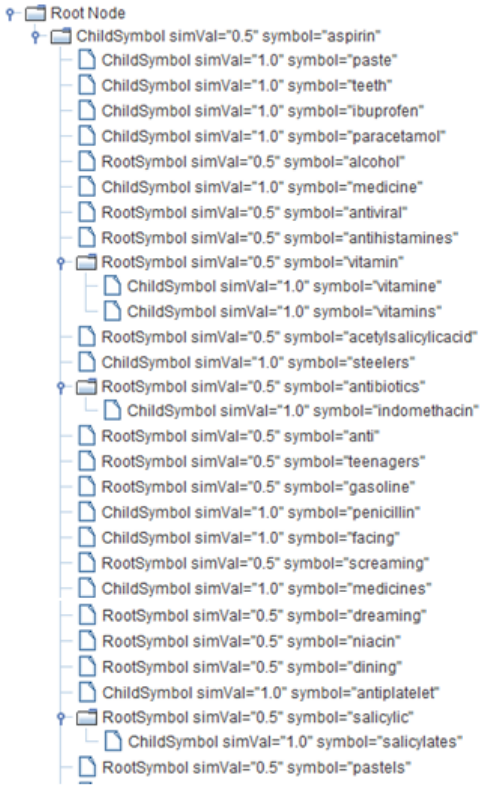

Fig. 6. Example taxonomy extracted for the search term 'Aspirin'

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. 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)