SWeMoF: A semantic framework to discover patterns in learning networks

PREPRINT! Please cite as: Kalz, M., Beekman, N., Karsten, A., Oudshoorn, D., van Rosmalen, P., van Bruggen, J., & Koper, R. (2009d). SWeMoF: A semantic framework to discover patterns in learning networks. In Cress, U., Dimitrova, V., & Specht, M. (Eds.). Learning in the Synergy of Multiple Disciplines. Proceedings of the Fourth European Conference on Technology-Enhanced Learning. Nice, France. Lecture Notes in Computer Science Vol. 5794. pp. 160-165. Berlin:Springer-Verlag.

SWeMoF: A semantic framework to discover patterns in learning networks

Marco Kalz[1], Niels Beekman[2], Anton Karsten[2], Diederik Oudshoorn[2], Peter Van Rosmalen[1], Jan Van Bruggen[1] and Rob Koper[1]

[1] Open University of the Netherlands, Center for Learning Sciences and Technologies, PO Box 2960, 6401 DL Heerlen, The Netherlands
[2]Open University of the Netherlands, Faculty of Informatics, PO Box 2960, 6401 DL Heerlen, The Netherlands

{marco.kalz, peter.vanrosmalen, jan.vanbruggen, rob.koper}@ou.nl
{cg.beekman, as.karsten, dj.oudshoorn}@studie.ou.nl

In this contribution we introduce SWeMoF, a semantic framework to discover patterns in learning networks and the blogosphere. Based on a description of the state of the art in data mining, text mining and blog mining we discuss the architecture of the Semantic Weblog Monitoring Framework (SWeMoF) and provide an outlook and an evaluation perspective for future research and development.

Keywords: weblogs, social software, text mining, data mining, RSS, clustering, classification, Latent Semantic Analysis

1 Introduction

In the past we have concentrated on the evaluation of Latent Semantic Analysis (LSA) to approximate the prior knowledge of learners in learning networks. We could show that Latent Semantic Analysis (LSA) is a promising method to support this process [1]. Several other examples show that semantic services and language technology have the potential to help to reduce tutor load and to increase efficiency in technology-enhanced learning [2]. We expect that the application of such approaches can help in personalization processes, the automatic generation of metadata and the discovery of structural patterns in learning networks. On the other hand we made the experience that the effort to develop and evaluate learner support services based on text- and data-mining methods is a very challenging task since a lot of different tools and sources are involved and manually processing of data is needed. Based on this aspect and the need to extend our research to other methods and approaches we have developed a prototypical solution that can help to find semantic patterns in learning

networks. In this contribution we present the Semantic Weblog Monitoring Framework (SWeMoF). The prototypical framework that we discuss in this article employs feed parsing techniques and data- and text-mining algorithms for several types of experiments and prototyping scenarios.

A similar framework as proposed here has been described by Joshi & Belsare as BlogHarvest [3] and Chau et al. [4]. The BlogHarvest framework is a conceptual framework for opinion and sentiment analysis that employs part-of-speech tagging, association rules and several miners for clustering and classification. The second proposal by Chau et al. consist of a blog spider to collect content, a blog parser to extract information, a blog analyzer and a blog visualizer. On the other hand this is a very general framework without any prototype or a detailed architecture.

The Semantic Weblog monitoring Framework (SWeMoF) enables researchers, course designers and learning technology developers to conduct several kinds of semantic experiments using different algorithms from natural language processing and data mining. We expect this framework to support the development of semantic technologies and web services to solve some basic problems in educational technology like formulated by Koper [5]. Applying common data and text mining techniques for discovery, recommendation and similarity classification can help to overcome problems of efficiency and effectiveness of the learning process and the workload of tutors. In the next part of the paper we describe the state-of-the art in data mining, text mining and blog mining. Afterwards we introduce the architecture of SWeMoF and provide an evaluation outlook.

2 Data Mining, Blog Mining and Text Mining

Data Mining is a process to find patterns in large numbers of data [6]. While data mining is applied most of the time to numerical data in large databases the application of techniques from data mining to textual data is called text mining. Inside the text mining research the application of text mining to weblogs is called blog mining.
The target of data mining is to discover meaning in a vast amount of data and to find patterns that are not recognizable by traditional statistical measurement and direct visual inspection. Witten and Frank refer to an increasing gap in today's society between the generation of data and the understanding of it [6]. In this sense data mining does not have the target to generate new data but to use existing data and to find structures which have not been explored before. Fayyad et al. describe the data mining process as an interactive and iterative process which involves several steps with different tasks [8]. A special focus on using data mining in educational settings and with educational data has been developed in the last years and applied to different educational problems [9] .

The same procedure as described above can also be applied to non-numeric data. If data mining is applied to text the process is called text data mining or text mining. Hearst defines text mining as a means of exploratory data analysis and he stresses the distinction between text mining and information retrieval [9]. While information retrieval and information access are only about finding information which are hard to find because of a lot of similar information text mining in his opinion is a process

which has the target to discover information that have never been encountered before. Several disciplines contribute to text mining research, the most important one computational linguistics/natural language processing (NLP) [10]. In addition several disciplines from literature studies to genetics and bio-informatics have applied text mining to solve some basic problems in their domain of research. The application of data mining techniques and text mining problems to weblogs is coined as Blog Mining. Blog mining is a very recent research direction. Barone provides a good overview of research done in this area until 2007 [11]. The framework proposed in this contribution will allow blog mining experiments with a special focus on discovery, classification and clustering. In the next part we describe the architecture of the framework.

3 Architecture of the Semantic Weblog Monitoring Framework

SWeMoF is an object-oriented, web-based application designed for semantic experiments on the basis of content produced from weblogs and other text-based applications which offer an RSS-feed. Within this framework several data mining/natural language processing experiments are possible. Every experiment takes the content of one or more weblogs as input, applies one or more algorithms/miners to the content and gives an output which can be downloaded. The level of input can be the whole content of a weblog (set level), content from a dedicated category in a weblog (category level) or even only dedicated postings (document level).

The prototype has implemented 5 example algorithms/miners for three different experiments: Semantic Similarity, Classification and Clustering. The prototype is written in Java and makes use of an integrated database and the Echo framework for the interface. The example algorithms are implemented using the Weka framework, but the SWeMoF framework does not depend on it. Both filters and text mining algorithms can be written from scratch or by using any available components and libraries. For the design of the system the following use cases have been defined:

- Corpus Creation

  A corpus has to be defined before an experiment can be created. This corpus can be constructed from several RSS-Feeds and/or OPML files. Besides this functionality, the domain corpus can be combined with a general language corpus which has been discussed as an important option in several information retrieval scenarios. For classification experiments several examples need to be classified manually before an experiment can be executed. In the classification experiment these ‘gold standard’ examples are needed to allow a semantic comparison between the classified documents and the unclassified documents. This step can be done by inspecting the corpus directly or during the creation of an experiment.

- Experiment Creation

  In the experiment creation phase the parameters for a text mining experiment can be configured. These parameters consist of a corpus, an optional general language corpus, filters and a text mining algorithm. Further, the level on which

the experiment is conducted (set, category or document) must be configured. It is also possible to disable a part of the corpus on any level: set, category or document. After an experiment has been created it can be executed. This division between experimentation and execution allows for repeating experiments and comparing results with different settings.

- Result Presentation & Download

After the execution of an experiment the results are presented to the user and the user can download the results.

- Adding of additional miners

In the current prototype the following miners have been implemented: Naive Bayes Classifier, IB1 Classifier, EM Clusterer, Simple K-Means Clusterer and a similarity rater using LSA. In addition, LSA can be combined with the miners implemented. But it is easy to add additional miners into the system.

The SWeMoF Framework allows the user to either create new experiments or retrieve and execute older experiments that have been stored. The parameters of an experiment (corpus, general language corpus, filters, miner, mining level) are saved in an experiment configuration. The following figure shows how a text mining experiment is conducted with SWeMoF.

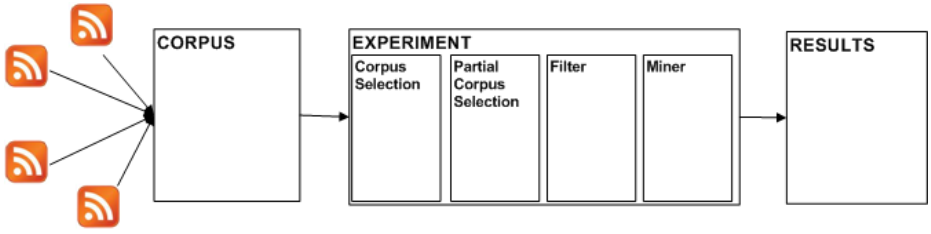

Fig.1. Overview of the components of the SWeMoF system

The *Input* module is responsible for the import of text. Single texts (e.g. single web posts) are organized in groups to create a hierarchy. Single text documents must be grouped in a document category, document categories must be grouped in document sets. Since SWeMoF's main focus is on web feeds, this design has been chosen to reflect the structure of these feeds. Even when only a single text document is imported it will have to be placed inside a document category, and the document category inside a document set. It is important to note that the corpus is not created by the Input module but by the Corpus module. For the feed parsing we have used the ROME library. ROME is a set of open source Java tools for parsing, generating and publishing RSS and Atom feeds. The *Corpus* module is responsible for the aggregation of documents generated from the input text by the input module. A corpus contains a collection of document sets. The structure within these sets is as described in the Input module section. After execution of an experiment the results are generated. The *View* module can display these results in different ways. In the prototype the *View* module is not implemented, instead a textual output is generated

directly from the *Result* object. Three types of information can be stored through the *DAO* module: the corpus, the configuration parameters of the experiment, and the results of an experiment. Finally a *GUI* takes care of the interaction with the user, enabling him to create new experiments, retrieve old experiments, retrieve results of experiments, and set parameters of an experiment. The *GUI* takes care of the interaction between users and the SWeMoF application. It is designed to let the user select text to convert (single document or web feeds); select filters (preprocessors) to generate the appropriate text corpus; select an experiment and choose a way to study the outcomes of the experiment. The GUI has been implemented with the Echo web framework.

The SWeMoF framework can be extended on several areas. The framework focuses on weblog monitoring and thus the focus for the prototype has been on implementing RSS and OPML as the document source. The input module however is designed in such a way that it can easily be extended with other input sources by implementing the appropriate interfaces. The second more important part where SWeMoF can be extended is in the filters and miners. To add a new filter or text mining algorithm all that needs to be done is implement the interface Filter or Miner and create a descriptor. The descriptor will tell the GUI what the Filter or Miner does and which options can be set. After this has been done, the descriptor can be added in to the registry. SWeMoF will then automatically make this filter or miner available to the end user.

4 Discussion, Outlook and Future Work

At the current stage of the development we could conduct several tests related to code functionality and result quality. After the components have been tested alone the integrated system has been tested to see if the system supports the use cases for which it was designed for. In addition we have compared the system results with the results of using Weka directly. The integration testing confirmed that the system is able to support the use cases and the comparison to Weka was successful as well. A real end-user and usability testing could not be conducted yet, but we are planning to present the system to researchers and learning technology developers with different levels of prior knowledge about data and text mining. For this purpose we are planning to combine traditional usability testing with the hedonic and pragmatic approach developed by Hassenzahl [13]. In this framework the “hedonic quality” aspect covers non-task-oriented quality aspects like innovativeness or originality and takes appealingness of a software system into account as well.

As a next step we will conduct an end-user testing with colleagues in the field. Based on the feedback of the potential end-users we will improve the system. The full code of the framework has been released under a GPL license [14] and a demonstration of the framework is available [15]. Depending on the reaction of end-users of the system we might improve the storage and presentation of the results. In addition we are going to extend the system with more miners from Weka and use it as an evaluation instrument for the development of several semantic web-services in the future.

Acknowledgements

The work presented was partially carried out in the by the TENCompetence Integrated Project that is funded by the European Commission's 6th Framework Programme, priority IST/Technology Enhanced Learning. Contract 027087 (www.tencompetence.org) and partially carried out as part of the partially carried out as part of the LTfLL project, which is funded by the European Commission (IST-2007-212578) (http://www.ltfll-project.org).

References

[1] M. Kalz, J. Van Bruggen, B. Giesbers, W. Waterink, J. Eshuis, & R. Koper (2009). Where am I? – An Empirical Study about Learner Placement based on Semantic Similarity. Manuscript submitted for publication.

[2] P. Van Rosmalen (2008). Supporting the tutor in the design and support of adaptive e-learning. Doctoral Dissertation. SIKS Dissertation Series 2008-07. Heerlen:Open University of the Netherlands.

[3] M. Joshi and N. Belsare: BlogHarvest: Blog mining and search framework, Proceedings of the 13th International Conference on Management of Data (COMAD), L.V. Lakshmanan, P. Roy, and A.K. Tung, Delhi, India, Computer Society of India, 2006.

[4] M. Chau, J. Xu, J. Cao, P. Lam, and B. Shiu: A Blog Mining Framework, IT Professional, vol. 11, 2009, pp. 36-41.

[5] R. Koper: Use of the semantic web to solve some basic problems in education: Increase Flexible, Distributed Lifelong Learning, Decrease Teacher's Workload., Journal of Interactive Media in Education, vol. 6, 2004, pp. 1-23.

[6] I. Witten and E. Frank: Data Mining. Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems):, Morgan Kaufmann, 2000.

[7] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy: Advances in knowledge discovery and data mining, AAAI Press, 1996.

[8] C. Romero and S. Ventura: Educational data mining: A survey from 1995 to 2005, Expert Systems with Applications, vol. 33, 2007, pp. 135-146.

[9] M. Hearst: Untangling text data mining, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics, 1999, pp. 3-10.

[10] C.D. Manning and H. Schutze: Foundations of Statistical Natural Language Processing, MIT Press, 2003.

[11] F. Barone: Current Approaches to Data Mining Blogs, Kent, University of Kent, 2007.

[12] C.H. Brooks and N. Montanez: Improved annotation of the blogosphere via autotagging and hierarchical clustering, Proceedings of the 15th international Conference on World Wide Web, Edinburgh, Scotland, ACM Press, 2006, pp. 625-632.

[13] M. Hassenzahl: The Effect of Perceived Hedonic Quality on Product Appealingness, International Journal of Human-Computer Interaction, vol. 13, 2001, pp. 481-499

[14] Semantic Weblog Monitoring Framework project page, http://swemof.sf.net

[15] Semantic Weblog Monitoring Framework demonstration page, http://www.swemof.org