Empowering web portal users with personalized text mining services

Fedor Bakalov[1]✉, Marie-Jean Meurs[2,3]✉, Birgitta König-Ries[1], Bahar Sateli[2], René Witte[2], Greg Butler[2,3], Adrian Tsang[3,4]

[1]Institute for Computer Science, Friedrich Schiller University of Jena, Jena, Germany
[2]Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
[3]Centre for Structural and Functional Genomics, Concordia University, Montreal, Canada
[4]Department of Biology, Concordia University, Montreal, Canada

Motivation and Objectives

Nowadays, many organizations use portals extensively as a single-point access to information, applications, and people. However, dealing with the constantly growing amounts of information available through web portals is difficult and time-consuming for users. Most of the current portal systems enable users to retrieve content statically defined as relevant - but reading and interpreting it remain a serious bottleneck. We propose to break this bottleneck with a personalized information system that integrates Natural Language Processing (NLP) to support users in analysing, transforming, and creating knowledge from large amounts of textual content. Our approach is a novel combination of web portal technology with the Semantic Assistants (SA) framework (Witte and Gitzinger, 2008), an extensible software architecture that allows invoking literally any NLP or text mining tool using either Web Services or Application Programming Interfaces (API). The whole system is designed to give users full control over personalization, and leverage visualizations to adjust the adaptive behaviour to the users' preferences in an easy-to-use way.

Methods

The proposed system relies on three major components: a web portal compliant with the Java Portlet Specification JSR286; the Semantic Assistants framework providing NLP services; and a module dedicated to user modelling and personalization. Web Portals are web applications providing users with unified access to various information resources and services. The most widely used industry standard for portal technology is the Java Portlet Specification JSR286. This standard defines an API for developing portlet applications in the Java programming language. A portlet is a pluggable user interface component that provides a specific piece of content or an application. Portlets can be aggregated into a portal page. Semantic Assistants are an existing open source service-oriented framework that brokers NLP pipelines as W3C standard web services. This framework brings NLP techniques directly to end users by integrating them within desktop applications. User Modelling and Personalization components allow storing information on user interests, which are represented as an overlay of domain concepts defined in the domain ontology. For each concept, the user model stores the exact degree to which the user is interested in it. The user model is updated following our hybrid approach (Bakalov et al., 2009). The portal content is delivered to users through personalizable portlets that can be viewed in standard or personalized states. In a personalized state, users can choose between several personalization effects (e.g., sort content by interest or chronologically).

Genozymes Portal: This biochemical literature portal has been developed for the Genozymes project at Concordia's Centre for Structural and Functional Genomics (CSFG) and is currently in use by a group of biologists, biochemists and geneticists working on lignocellulose research. The goal of this research is to find novel ways of creating bioproducts and biofuels from green waste. Part of this work is the curation of content regarding specific enzymes of fungal origin from the domain literature. Towards this end, literature from the PubMed portal needs to be evaluated for relevance, which is a time-consuming task. To support these researchers, we automatically import new articles appearing on PubMed into a portal (Figure 1), processing them with the myCoMINE NLP pipeline (Meurs et al, 2012), which extracts entities and facts related to fungal enzymes. The Query portlet displays user's search queries. These queries can be hierarchically organized and modified by adding, renaming or deleting keywords. The Listing portlet presents the most relevant papers found among new articles appearing on PubMed with regards to all or a selected subset of the user queries. In our example (Figure 1), the mention of cellulose percentage

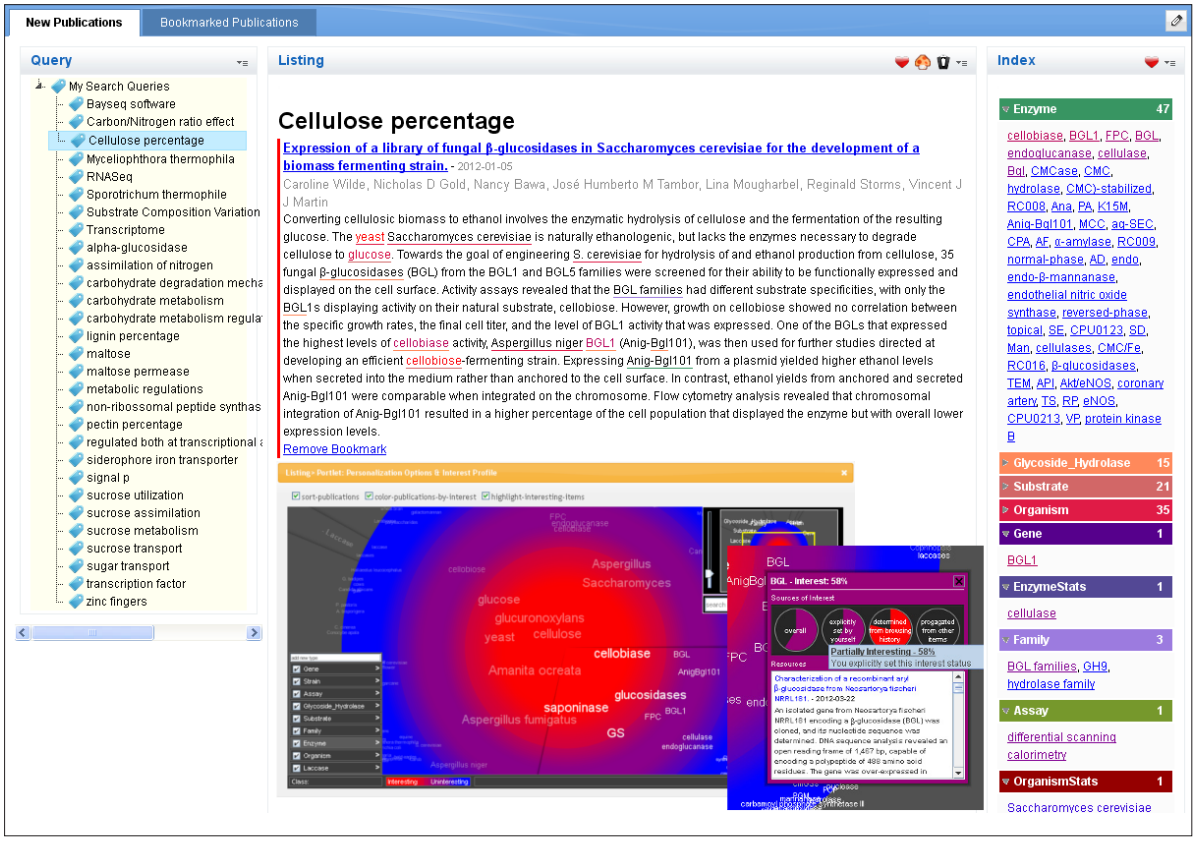

Figure 1: Genozymes portal with IntrospectiveViews

has been selected in the Query portlet and the user has requested the mycoMINE assistant on the papers appearing in the Listing portlet. The Index portlet displays the mycoMINE results in terms of entities and facts mentioned in the papers. The number of different occurrences is indicated for each type of entity and fact (e.g. 47 different enzymes). Each reported entity or fact is linked in the texts to their corresponding mentions, which are underlined, then highlighted when the user selects them in the Index portlet. Portal content and SA results can be personalized by users. In the personalized view, users can view and edit their interest profile as well as define how the portlet content should be personalized. This is done in a personalization options window (Figure 1 - bottom) displayed as an overlay over the portlet. The personalization options vary from portlet to portlet. For example, the Listing portlet supports three personalization effects which can be selected by checking the corresponding checkboxes: (1) sorting publications according to the user interest profile; (2) highlighting the most interesting of the user publications by a colour marker; (3) highlighting mentions of items from the user interest profile in the publications list. User changes on the personalization options are immediately projected onto the portlet content. The personalization interface we proposed (Bakalov et al., 2010) visualizes user interests using a metaphor of circular zones partitioned into slices, where each zone represents items of certain interest degree and each slice represents items of a specific type. The hot zone in the centre displays items that users are strongly interested in, while the cold zone at the circle edge displays less interesting items. The visualization follows Shneiderman's information seeking mantra (Shneiderman, 1996) by providing functions for getting an overview, zooming in and out, filtering, searching and giving detailed information about items upon request (Figure 1 - bottom, details). The visualization also allows editing information in the model (adding and deleting items, changing interest degree, etc.). Similar to the changes of personalization options, all changes in the interest profile made through the visualization are immediately projected on the personalized con-

tent. For example, upon a change in the interest profile, the publications in the Listing portlet will be re-sorted and the colour markers of the most relevant publications will be updated.

Results and Discussion

To evaluate the impact of introducing personalized text mining services in our Genozymes portal, we conducted a user study with seven CSFG researchers. The results of this evaluation showed that providing users control over personalization and text mining services makes substantial impacts on the usefulness, usability, and user satisfaction of the personalized system. The Genozymes portal is available for demonstration at http://www.minerva-portals.de:10040/wps/portal, user=demo, pswd=portaluser.

We demonstrate how to enhance web portals with personalized text mining services that enable users to focus on the interesting sections of the presented contents. In such a way, portal users can apply NLP tools not only on publications, but also on a variety of other resources and applications that can be aggregated using portal technology, such as patents, databases, samples, or observation and sensor data.

Acknowledgements

We thank all the participants of the user study for their contribution. We thank Justin Powlowski for his expert advice on the biology content. Funding for part of this work was provided by Genome Canada and Genome Quebec. Part of this research was sponsored by the IBM Ph.D. Fellowship Awards Program and carried out in the framework of the Minerva Portals project in cooperation with IBM Deutschland Research & Development GmbH.

References

1. Bakalov F, König-Ries B, et al. (2009) Hybrid Approach to Identifying User Interests in Web Portals. Int. Conf. on Innovative Internet Community Systems, 2009.
2. Bakalov F, König-Ries B, et al. (2010) IntrospectiveViews: An interface for scrutinizing semantic user models, Int. Conf. on User Modeling, Adaptation, and Personalization.
3. Meurs MJ, Murphy C, et al., (2012) Semantic text mining support for lignocellulose research, BMC Medical Informatics and Decision Making 12(Suppl 1):S5. doi:10.1186/1472-6947-12-S1-S5
4. Shneiderman B (1996) The eyes have it: A task by data type taxonomy for information visualizations. IEEE Symposium on Visual Languages, 1996.
5. Witte R and Gitzinger T (2008) Semantic Assistants - User-Centric Natural Language Processing Services for Desktop Clients. Asian Semantic Web Conference, LNCS5367-360.