

DOMAIN MODELLING IN SOCIALLY AWARE INFORMATION SPACES

Marián ŠIMKO

*Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
marian.simko@stuba.sk*

Abstract. Semantic descriptions about domain resources are necessary to enable machines to understand resources' content and adopt the content for a user. We provide an overview of our approaches to automated domain model acquisition by processing different sources of information. We particularly focus on socially aware information spaces and take advantage of the content provided by users themselves.

1. Domain modelling

When accessing information present in open or closed information spaces, researchers may face well-known information overload problem. It is impossible to absorb vast amounts of research and other types of documents. The need for intelligent processing of the content such as recommendation or filtering is inevitable. To enable intelligent processing, the knowledge about content being processed – known as domain semantics – is necessary.

Forms of semantics to describe information resources vary in expressivity and formality [9]. Although advanced, heavyweight semantics is desirable, it is often difficult to provide it. Manual creation of complex semantic structures such as domain ontologies is demanding and difficult task for a human. Especially when considering large information spaces, which change dynamically as a result of social activity and interactions with the content particularly popular for the last decade. Automatic methods for complex semantics extraction do not succeed to a sufficient extent. Therefore, initiatives promoting the idea of lightweight semantic descriptions emerge [6, 7, 8]. Lightweight semantics suffices for the majority of basic information processing tasks, constituting an appealing trade-off between complexity of automatic creation and “performance” of semantic representation.

Besides being a dynamically changing environment, socially aware information spaces involve big amounts user annotations, which represent a challenging source of addi-

tional information to be processed when automatically creating or enriching domain semantics.

2. Our approaches to domain model acquisition

In our work we have focused on different sources of information to process. We tackled user-created content in different settings to utilize it for producing domain semantics (referred to as a domain model) in an automated way (see Table 1).

Table 1. Overview of approaches to automated domain model acquisition.

Approach	Major focus of approach	Dominant information source to process	Application domain type (open-/close-information space)
Majer, Šimko [3]	<i>Resource relevance identification</i>	<i>user-created annotations (microblog posts)</i>	<i>open</i>
Uherčík, Šimko, Bieliková [5]	<i>RDT extraction</i>	<i>user-created annotations (microblog posts)</i>	<i>open</i>
Lučanský, Šimko [2]	<i>RDT extraction</i>	<i>document visual style</i>	<i>open</i>
Harinek, Šimko [1]	<i>RDT extraction</i>	<i>user-created annotations (tags, highlights, comments)</i>	<i>close</i>
Svrček, Šimko [5]	<i>RDT description enrichment</i>	<i>user-created annotations (definitions)</i>	<i>close</i>

In [3] we presented a method for ranking resources referenced by microblog users in their posts. We view microblog posts as user-created annotations of the referenced resources. The resource ranking is computed by leveraging microblog social network structure. The contribution of this work lies in distinguishing between resources that are less or more relevant, i.e., less or more suitable sources for information to mine.

Microblog posts were considered as annotation of web content in [5] as well. The proposed method for relevant domain terms (RDT) extraction utilizes social characteristics of the Twitter network to derive relevancies of microblog posts assigned to the web resources. Microblog posts' relevancies affect final weights of extracted RDTs. The experiments showed that microblog data can facilitate domain model basic descriptions creation.

In [2] we proposed a method for relevant domain terms extraction from web pages combining traditional automatic term recognition algorithms with web site's visual style processing, while particularly focusing on cascade style sheets. The proposed approach confirmed the visual style provided by web page creators as important factor to consider when deciding about relevance of terms on the page as of semantic descriptors.

In [1] we presented the method for extracting relevant domain terms from documents having assigned various types of user-created annotations. The novelty of our approach lies in exploring the contribution of various annotation types to term extraction. We considered three very elemental annotation types: tags, comments and highlights. The method was evaluated in educational domain. The result showed that different annotation types processing contribute to overall semantics with different relevance and that combination of annotation types surpasses the partial results.

We also presented an experiment with educational activity of definition creation in a learning system [5]. The "crowd-sourced" output of the activity can be utilized for domain

model enrichment: filtered student-created definitions can be mapped to domain model's RDTs and provide an intentional definition of concepts represented by those RDTs.

Each our approach represents an advance in automated domain modelling, either in the pre-processing step [3] or in the phase of selected domain model part creation or enrichment [1, 2, 4, 5].

Acknowledgement: This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11, the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010, the Slovak Research and Development Agency, contract No. APVV-0208-10, and it is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF, the Research and Development Operational Program for the projects Support of Center of Excellence for Smart Technologies, Systems and Services SMART, ITMS 26240120005 and SMART II, ITMS 26240120029, co-funded by ERDF.

References

to other papers publishing the results that are summarized here

- [1] Harinek, J., Šimko, M. Improving term extraction by utilizing user annotations. In *Proc. of 13th ACM Symposium on Document Engineering*, ACM, (2013), pp. 185–188.
- [2] Lučanský, M., Šimko, M. Improving Relevance of Keyword Extraction from the Web Utilizing Visual Style Information. In *Proc. of 39th Conference on Current Trends in Theory and Practice of Computer Science, SOFSEM 2013*, Springer, (2013), pp. 445–456.
- [3] Majer, T., Šimko, M. Leveraging Microblogs for Resource Ranking. In *Proc. of 38th Conference on Current Trends in Theory and Practice of Computer Science, SOFSEM 2012*. Springer, (2012), pp. 518–529.
- [4] Svrček, M., Šimko, M. Supporting Educational Content Enrichment and Learning via Student-Created Definitions. In *7th International Workshop on Social and Personal Computing for Web-Supported Learning Communities, SPeL, at ICWL 2014*. To appear, Springer.
- [5] Uherčík, T., Šimko, M., Bielíková, M.: Utilizing Microblogs for Web Page Relevant Term Acquisition. In *Proc. of 39th Conference on Current Trends in Theory and Practice of Computer Science, SOFSEM 2013*. Springer, (2013) pp. 457–468.

Other references

- [6] Barla, M. Towards Social-based User Modeling and Personalization. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, (2011), Vol. 3, No. 1, pp. 52–60.
- [7] Bielíková, M., Barla, M., Šimko, M. Lightweight Semantics for the “Wild Web”. In B. White, P. Isaias and F. M. Santoro (eds.): *Proc. of the IADIS Int. Conf. on WWW/Internet*, ICWI 2011, IADIS Press, (2011), pp. xxv–xxxii.
- [8] Giunchiglia, F., Zaihrayeu, I. Lightweight ontologies. Technical Report DIT-07-071, University of Trento, (2007).
- [9] Wong, W., Liu, W., Bennamoun, M. Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR)*, ACM, (2012), 44(4), no. 20.