LIGHTWEIGHT SEMANTICS FOR THE “WILD WEB”

Mária Bieliková, Michal Barla and Marián Šimko
Slovak University of Technology in Bratislava, Faculty of Informatics and Information Technologies,
Institute of Informatics and Software Engineering, Ilkovičova 3, 842 16 Bratislava, Slovakia
{bielik, barla, simko}@fiit.stuba.sk

ABSTRACT

The current Web has many aspects. It is no longer only a place for content presentation. The Web is more and more a place where we actually spend time performing various tasks, a place where we look for interesting information based on discussions, opinions of others, as well as a place where we spend part of our recreation and leisure time. In addition, the Web provides an infrastructure for applications that offer various services. In this paper we concentrate on representation and acquisition of lightweight semantics for the “wild” Web, which is a must if we want to shift to a “smarter” Web and web applications, which cope with dynamic content and take into account user features to deliver personalized experience. We already have mechanisms that infer recommendations for a particular user where the context is known and the content is described using a particular form of semantics (which is a case of only a few islands in the vast ocean of the Web). In moving to the “wild Web” we do not have many clues about the content itself. More often we have a picture of activities performed within particular content, which can help at least as well as the content itself. We present our proposal of lightweight semantics models together with their social enhancements and discuss some aspects of lightweight semantics acquisition on the “wild Web” as large and dynamic information space. We discuss examples of approaches to towards an improvement of fulfilling our information needs based on reasoning on semantic description of the web content. These examples are recent results originated from the PeWe (Personalized Web, pewe.fiit.stuba.sk) research group at the Institute of Informatics and Software Engineering at the Slovak University of Technology in Bratislava.

KEYWORDS

Lightweight semantics, domain model, user model, annotation, adaptive proxy, relevant domain term

1. INTRODUCTION

A requirement for having an additional description of the Web content, for knowing its semantics and thus allowing machine processing, is almost as old as the Web is. A good example are classical search engines, which were relying on metadata tags manually inserted into static web pages in order to respond better to user queries.

While the semantics has been a primary focus for the Semantic Web initiative (Berners-Lee, 1999), it is crucial also for the typical “wild” Web of nowadays, which provides far more than a static content – it has become highly dynamic, it provides *functionality* on the top of the content and due to large amount of information it is becoming more and more *personalized* as we got used to *use* the Web through its services as our primary source of information and knowledge but also as a mean to connect with our friends or other people of our interest.

All this makes the requirement for semantics even more important – if we want to build the services which tailor their functionality and underlying content to the needs of a particular visitor, we need to know not only about features of that visitor, but also about the content.

Any information space, including the Web, can be viewed as a large semiotic system consisting of symbols that represent world around us according to some conventions. The idea is related to semiotic triangle principle (Frege, 1892) that distinguishes between *concept* – abstract thought, *object* – physical realization of that concept, and *symbolic* representation used in a natural language (see Figure 1). A relationship between an object and a concept originates in cognition and is direct. It is referred to as designation. A relationship between a concept and a symbol refers to symbolic representation of a concept and it is referred to as signification. A relationship between a symbol and an object is imputed and it is referred to as denotation. A concept is the mediator that relates symbol to its object. A symbolic representation can be viewed as both interpretation of concept and reference to an object.

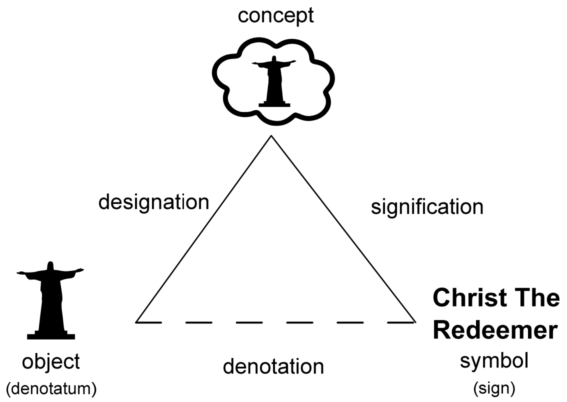

Figure 1. Semiotic triangle (according (Frege, 1892)).

When providing semantics for web content, we interpret symbols and associate them with an appropriate sense in form of conceptual description. The complexity of descriptions may vary. Term *ontology* is used to represent knowledge about a particular domain or among domains. The widely accepted definition by Gruber states that ontology is an explicit formal specification of a shared conceptualization of some domain (Gruber, 1993). However, the notion of ontology differs among researchers based on a degree of formality and expressiveness, or a possibility to specify axiom schemas and general axiomatic theorems for reasoning. We started to differentiate more frequently lightweight and heavyweight ontologies (Uschold, Gruninger, 2004; Giunchiglia, Zaihrayeu, 2007; Wong et al., 2011).

Where heavyweight ontologies contain advanced structures such as axioms and enable complex reasoning, lightweight ontologies form only basic conceptual structures. For example, Giunchiglia and Zaihrayeu define lightweight ontology as follows (Giunchiglia, Zaihrayeu, 2007):

> “A (formal) lightweight ontology is a triple $O = \langle N, E, C \rangle$, where N is a finite set of nodes, E is a set of edges on N, such that $\langle N, E \rangle$ is a rooted tree, and C is a finite set of concepts expressed in a formal language F, such that for any node $n_i \in N$, there is one and only one concept $c_i \in C$, and, if n_i is the parent node for n_j, then $c_j \sqsubseteq c_i$.”

Concept definition in lightweight ontologies is slightly simplified and “lightweight concepts” often represent rather a lexical reference to concepts than concepts themselves. Examples of semantic representation falling into a notion of lightweight ontology involve (Wong et al., 2011):

- terms,
- glossaries/dictionaries,
- thesauri,
- taxonomy.

A specific form of lightweight domain conceptualization is folksonomy. With the emergence of Web 2.0 a plethora of users participate in web content creation and enrichment as they share, organize, manage and annotate resources, which they access. The most popular form of user contribution to the web content is tagging, process of assigning hand-picked terms to web resources. Resources with tags associated by many users form a basis for deriving a folksonomy out of it. Tags are very similar to keywords being assigned to a web page by the page creator, but they represent rather objective notation of page content as they originate from different users. Tags are considered to be coarse-grained and informal, but they are also more accessible to a human user (Wu et al., 2006) and since users tend to reach a common agreement on domain vocabulary used for tagging, resulting tags reflect quite accurately the web resources’ meaning. Folksonomy is a result of social cognitive consensus about certain domain. The potential of domain modeling leveraging user-generated content is emphasized by recent advances in tag relationships discovery, which shift a folksonomy beyond flat and implicit structure (Heymann, Garcia-Molina, 2005; Barla, Bieliková, 2009).

Heavyweight semantic representation of the Web is fostered by the Semantic Web initiative. Its aim is to provide web resources with machine-friendly descriptions. However, the vision of the Semantic Web is not fulfilled at a pace we wish. Among other findings it has been reported that (Sabou, et. al, 2007):

- existing semantic systems are restricted to a limited set of domains,
- the overall Semantic Web does not adequately cover specific terminology,
- many online ontologies have weak internal structure and
- online ontologies contain modeling errors and contradictory information.

A partial help for domain experts in the process of building ontology constitute methods for ontology learning, which are typically based on large corpora processing (Cimiano, 2006). However, the accuracy of state-of-the-art ontology learning tasks based on content processing is still not satisfactory (Wong et al., 2011). This especially applies for more complex tasks such as non-hierarchical relationship discovery or axiom acquisition, which produce semantics typical for heavyweight ontologies. Moreover, we should consider inherent properties of the “wild” Web such as its dynamics and openness. In essence it is impossible to acquire semantics manually except a small island of closed worlds managed by particular applications.

2. LIGHTWEIGHT SEMANTICS

We proposed lightweight semantics for modeling (both domain and user) open and dynamic information spaces. Here our main concern is automatic or at least semiautomatic acquisition of semantics. So the question is not what we can do with the semantics when it is perfect (in sense of its formality and expressiveness), but how to acquire it for constantly changing and in advance unknown content.

Our models consist of two layers: a *designate* layer and a *metadata* layer (see Figure 2). Designate layer covers resource and user abstractions. We distinguish dual representation of resources: resource instances and resource designates. Resource instances are low-level representation of web resources (e.g., the content represented using XML, HTML). Resource designates constitute resource abstractions residing in a model. They are associated with metadata about resources. Differentiation between instances and designates supports the notion of reusability and extendibility in terms of content resource’s lower level representation.

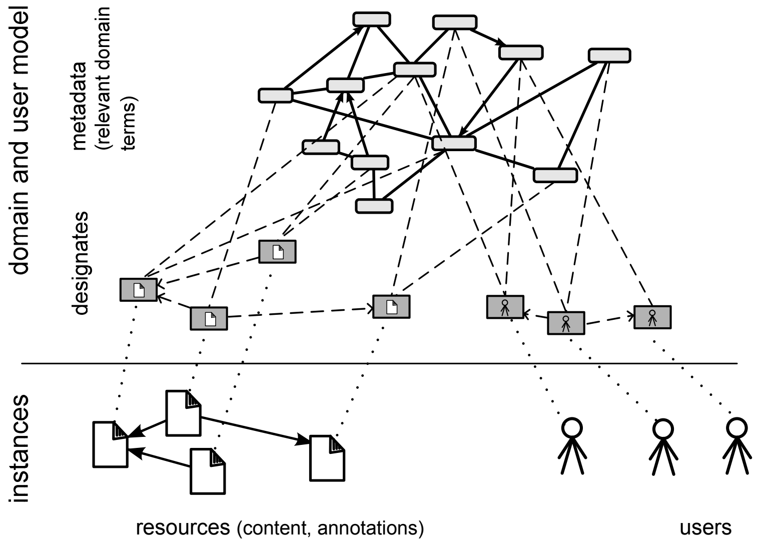
domain and user model
metadata (relevant domain terms)
designates
instances
resources (content, annotations)
users

Figure 2. Domain and user model represented using the same lightweight approach.

Users are represented by user instances – abstract entities that are detached from any conceptualization. User designates constitute abstraction over user instances aggregating users’ digital footprints in relation with their interaction in web environment (e.g., visited pages and pages’ representation).

Metadata layer is formed by relevant domain terms – easily creatable descriptions that are related to particular topics present in the content. It is important to note that relevant domain terms do not represent concepts in strict ontological definition, cf. (Cimiano, 2006). They rather represent lexical reference to the concepts, which form the models (unlike relevant domain terms, concepts are not explicit).

Elements in our models are interconnected via various types of relationships that represent various forms of relatedness between domain and user model elements. There are three high level types of entity to entity relationships between domain and user model elements:

- relationships between resource/user designates,
- relationships between resource/user designates and relevant domain terms,
- relationships between relevant domain terms.

Relationships between resource designates typically reflect relationships between resource instances such as hypertext links, subsumption relationships representing hierarchical book-like structure of a content or, friendship or similarity among users. Relationships between resource designates and relevant domain terms represent lightweight semantic descriptions of resources. Relationships between user designates and relevant domain terms represent user model, i.e., particular user characteristics related with the relevant domain term (e.g., interest). Using relationships between relevant domain terms we arrange relevant domain terms in a conceptual structure, which represents actual semantics used for reasoning on the content.

The proposed domain model is restricted neither to predefined types of domain elements nor relationship between them. The goal of multilayer design is to separate metadata and resources, which reflects into ability to easily define new types of entities that makes the models extendable and suitable to cope with dynamic changes typical for the Web 2.0.

The notion of metadata in the proposed models is slightly simplified (based on term-based descriptions) in order to achieve the degree of complexity, which facilitates domain model automated construction, while providing a solid basis for reasoning resulting in advanced functionality such as personalization.

3. ACTIVITY FLOWS IN THE WEB AND SOCIALLY-ENHANCED LIGHTWEIGHT SEMANTICS

Our lightweight semantics considers conceptualization that is created automatically or with the help of a small group of domain experts in case of closed “islands” of the Web information space. New activities and sources of semantics on the Web emerged as the Web reached the stage “2.0”. The activities on the current Web are not limited to “consumption” of information presented on static pages. Users do *not only* retrieve information, they also create own content, share and collaborate, communicate and discuss with others. The traditional activity of retrieving (useful) information is often accompanied or even replaced with activities supporting our current task or fulfilling our other needs, e.g., communication using chat or organizing resources by means of tags.

When considering activities that we perform, it is important to conceptually differentiate between their relation to the goal we try to achieve when accessing the Web. According to this relation, we divide activities into (Figure 3) core, and supportive. A core activity is an activity, which directly concerns our primary goal and purpose for accessing the Web and a particular site on it. The typical core activities are *searching* (either navigational or exploratory), *learning or shopping, depending on a type of visited web site.*

The core activity flow (solid line, Figure 3) covers the entire process of travelling through the Web in order to reach particular information. In this process, a variety of services are involved. In general, information processor services provide basic services, which process information according to particular goal. Users interact with the content and their actions are observed and tracked by the semantic logger. User model inferencer updates the user model by processing user actions captured in semantic logs.

Supportive activities are not directly related to a primary goal of the site visit, but are performed along with core activities in order to achieve primary goal faster and more efficiently. They often benefit from Web 2.0 enhanced user-centric interfaces. Supportive activities cover content tagging, providing descriptive annotations, participating in discussions or in other forms of collaboration. They are also related to different forms of resource rating.

The supportive activity flow (dashed line, Figure 3) covers activities related to web content enrichment by web users themselves. Actual web content presentation is obtained from the presenter. By using collaborative content creators users enrich the content. The content is created with respect to the current user model. Vice versa, performing an action related to the content creation reflects into the user model update.

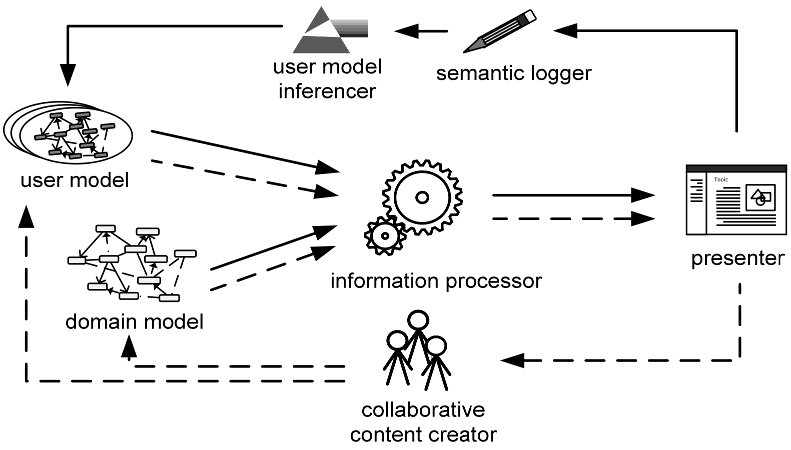

Figure 3. Activity flows in the Web 2.0: core flow: solid line, supportive flow: dashed line.

Both core and supportive activities performed by users can be in general viewed as an interaction between users and web resources described by metadata (see Figure 4). Supportive activities produce a new type of resource element, which is capable to enhance the conceptual structure representing lightweight semantics: a user-generated *annotation*. The annotation we see as fundamental mean that allows organize resources (e.g. using tags), create and share content (e.g. posting a comment), or interact (e.g. discussing particular topic). Annotation types added by the users vary depending on particular instance of collaborative content creator involved: tags, comments or highlighted text. Annotations are created by different users and it is important to track authorship of annotations as their quality or usefulness may differ.

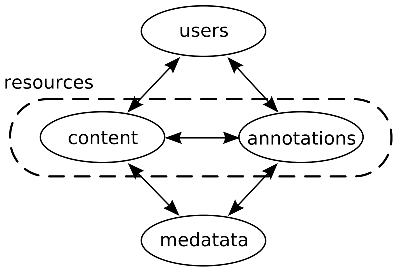

Figure 4. An abstraction of Web 2.0 core concepts: users create and access resources described by metadata.

Considering annotations as important source for the content description, i.e. metadata, we enhance our lightweight semantics model in two directions. We add:

- new resource type – *annotation* – to resource instances and resource designates as an important direct source for metadata,
- new element to resources – *creator* together with its abstraction on designates level as an indirect source for metadata.

Creator can be a reader who assigns a tag to a web page, or an author who adds relevant domain terms to the conceptualization. Furthermore, creator can be an artificial creator, e.g. a machine, which discovers relationship between relevant domain terms. The creator designate entities are important part of the model as they enable to recognize quantitative characteristics of resources produced by different authors (e.g. by users on different level of domain expertise or a generation method used with specified confidence of correctness).

4. AUTOMATIC SEMANTICS ACQUISITION FOR THE “WILD” WEB

We have two main options when considering acquisition of descriptions for the content of the “wild” Web: to base our approaches on available content and process it to extract required metadata, and/or to rely on users themselves to provide us with an evaluation of the content, its quality and usefulness for user’s current task. This evaluation is implicit, based on analysis of digital traces of user behavior within an information space.

Both of the mentioned approaches are useful if we want to employ lightweight modeling approaches for the purpose of information processing (including personalization, intelligent search) in the "wild" information space such as the Web is. We need an ability to acquire relevant domain terms from resources such as documents visited by the users and build the domain and user models on top of them. Because the Web is an open information space, we need to track down and process every page a user has visited in order to update her model appropriately. Apart from relevant domain terms, we need to acquire additional attributes describing the user's access to the web resource – implicit feedback indicators such as time spent actively reading a page or amount of scrolling.

To achieve this, we developed an enhanced proxy server, which allows for realization of advanced operations on the top of requests flowing from user and responses coming back from the web servers all over the Internet (Barla, Bieliková, 2010). Figure 5 depicts the user modeling flow supported by our platform. When a web server returns a web page as a response for a user's request, the proxy injects a specialized tracking javascript into it and passes the page to the client user agent. At the same time, it initializes a process of metadata extraction from the acquired page.

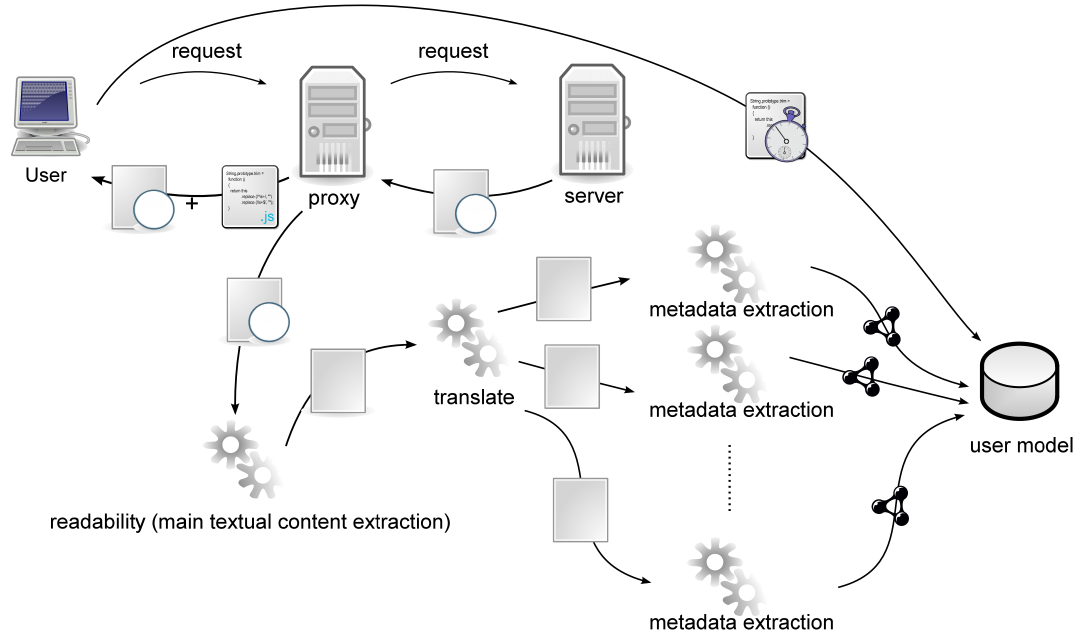

Figure 5. User Modeling process based on an enhanced proxy platform.

First, HTML page is processed by a *readability* module[1] which strips-off the HTML markup and leaves only a main textual content, omitting navigational parts, banners etc. Second, the text is translated into English (if it is not already in this language) using Google's translate service. This step is required as majority of metadata extraction algorithms and services (which are fired after the translation) work correctly only with English language. Extracted metadata are stored in a user model altogether with corresponding URL and timestamp. The tracking javascript, which was inserted to the response and passed to the user, supplies additional information about the user's activity within the page and thus adds an implicit feedback which determines a weight of contribution of just-discovered metadata to the whole user model.

The actual process of relevant domain terms extraction is based on both traditional natural language processing methods as well as on new online services such as OpenCalais (opencalais.com) or Alchemy (alchemyapi.com). The latter approach has a great potential as the services return not only relevant terms, but often provide also additional metadata including a binding to the concepts of LinkedData cloud. This allows us to disambiguate meaning of homonyms and establish relationships between relevant domain terms of our lightweight model. We believe that such a bottom-up approach for incorporating semantics is more viable than approaches relying on highly formalized ontologies.

[1] reimplemented solution available at http://lab.arc90.com/experiments/readability/

Besides relevant domain term identification also discovery of relationships between relevant domain terms is an important step in semantics acquisition process. Our aim is provide relationships in terms of lightweight semantics. We are particularly focused on most elemental relationships, which we believe are sufficient for multitude of tasks related to intelligent/advanced information processing. In particular, we consider *relatedness* relationship as basic form of paradigmatic similarity between terms, and *is-a* relationships that form hierarchical skeleton of domain conceptualization. We already showed that such approach is, despite the pitfalls associated with natural language processing, feasible in the domain of web-based learning (Šimko, Bieliková, 2009; Šimko, 2011).

Implicit feedback indicators, acquired by tracking javascript, inserted dynamically by our proxy, provide us not only with information on relevance of the user's particular access to the web content, but are also a basis for overall evaluation of the content, with respect either to all users of the system or to some selected virtual community.

5. CONCLUSIONS

The Web became a dynamic and constantly changing place. The emergence of Read/Write Web opened the web content to millions of users to collaboratively edit and organize it. At the same time, we need description of the content in order to provide advanced functionality such as smart and personalized search.

We present lightweight semantics modeling as an approach to address current challenges of the information processing on the "wild Web". We believe that lightweight semantics approach brings a benefit of a feasible automatic acquisition of semantic descriptions, while still being sufficient for majority of advanced information processing tasks. The major features of our approach related to both domain and user modeling for advanced/intelligent information processing are:

- Separation between domain conceptualization and content – the content and its metadata are separated in order to allow proper reusability of content (with no need to change domain conceptualization) and flexible information processing (metadata-based rather than content-based).
- Extendibility together with a possibility of new types of content – domain model facilitates definition and creation of new types of resources. This allows us to consider various types of web content that can be involved in advanced processing providing more functionality. User experience is increased.
- Reusability across various applications – a distributed nature of the Web resulted in various applications processing the web content. In order to shift advanced processing beyond one application, domain and user models are reusable across the Web supported by proxy.
- Explicit support for collaboration – interactivity and collaboration improve user experience and increases users' competences, which makes travelling through the Web more convenient.

Presented lightweight semantics models have been conceived as a result of recent research aimed towards an improvement of information retrieval and navigation within the Institute of Informatics and Software Engineering at the Slovak University of Technology in Bratislava, especially within the PeWe (Personalized Web) research group. Our results have been evaluated within several domains: digital libraries, news, learning and general web content (Bieliková et al., 2011). They include:

- acquiring a lightweight network of related terms and web objects annotations via the games with a purpose (Šimko, Tvarožek, Bieliková, 2011),
- automated creation of a domain model, mainly relationships discovery based on statistical and linguistic processing for adaptive educational learning framework ALEF (Šimko et al., 2011) and by leveraging collective wisdom of masses present in data of social tagging (Barla, Bieliková, 2009),
- collaborative news filtering based on generic full text engine exploiting power-law distributions (Suchal, Návrat, 2010) and content-based news filtering based on efficient vector and balanced tree representations (Bieliková, Kompan, Zeleník, 2011),
- faceted semantic exploratory browser providing adaptive views and personalized visual query construction, which works on semantically enriched information space (Tvarožek, 2011),
- query expansion by social context defined by a social network built from the stream of user's activity on the Web (Kramár et al., 2010),
- adaptive link recommendation based on an analysis of the user navigational patterns and his behavior on the web pages while browsing through a web portal (Holub, Bieliková, 2011).

ACKNOWLEDGEMENT

This work was partially supported by the grants VG1/0675/11/2011-2014, KEGA 345-032STU-4/2010, 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 authors wish to thank colleagues from the Institute of Informatics and Software Engineering and all students (members of PeWe group, pewe.fiit.stuba.sk) for their invaluable contribution to the work presented in this invited lecture.

REFERENCES

Barla, M., Bieliková, M. 2009. On Deriving Tagsonomies: Keyword Relations Coming from Crowd. *LNCS 5759: Proc. of the 1st Int. Conf. On Computational Collective Intelligence*, ICCCI 2009. Springer, pp. 309–320.

Barla, M., Bieliková, M. 2010. Ordinary Web Pages as a Source for Metadata Acquisition for Open Corpus User Modeling. In *Proc. of IADIS WWW/Internet 2010*. IADIS Press. pp. 227–233.

Berners-Lee, Tim; Mark Fischetti (1999). Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor. Britain: Orion Business. ISBN 0-7528-2090-7.

Bieliková, M., Kompan, M., Zeleník, D. 2011. Effective Hierarchical Vector-based News Representation for Personalized Recommendation. In *ComSis Journal.* To appear.

Bieliková, M., Návrat, P., Barla, M., Tvarožek, J., Tvarožek, M. (Eds.). 2011. Personalized Web – Sciences, Technologies and Engineering. *Proc. of 9th Spring PeWe Workshop,* Viničné, Slovakia, 110 p.

Cimiano, P. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag, ISBN: 978-0-387-30632-2. 347p.

Frege, G., 1892. Über Sinn und Bedeutung. *Zeitschrift für Philosophic und philosophische Kritik*, Vol. 100, pp. 25–50.

Giunchiglia, F., Zaihrayeu, I. 2007. Lightweight ontologies. Technical Report DIT-07-071, University of Trento.

Gruber, T. R., 1993. A translation approach to portable ontology specifications. Knowledge Acquisition. Vol. 5, No. 2, pp. 199–220.

Heymann, P., Garcia-Molina. H. 2006. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. *Infolab technical report*, Stanford. Available on: http://ilpubs.stanford.edu:8090/775/1/2006-10.pdf.

Holub, M., Bieliková, M. 2011. An Inquiry into the Utilization of Behavior of Users in Personalized Web. *Journal of Universal Science.* To appear.

Kramár, T., Barla, M., Bieliková, M. 2010. Disambiguating search by leveraging a social context based on the stream of users activity. *LNCS 6075, UMAP 2010*, Springer, 2010, pp. 387-392.

Sabou, M. et al. 2007. Evaluating the semantic web: a task-based approach. In International Semantic Web Conference, ISWC 2007, LNCS 4825, pages 423–437, Berlin, Heidelberg. Springer.

Šimko, J., Tvarožek, M., Bieliková, M. 2011. Little Search Game: Term Network Acquisition via a Human Computation Game. In *Proc. of HT 2011: 22nd ACM Conf. on Hypertext and Hypermedia*, New York: ACM, pp.57-61.

Šimko, M., Bieliková, M. 2009. Automatic Concept Relationships Discovery for an Adaptive E-course. In Barnes, T. et al. (Eds.). *Proc. of Educ. Data Mining 2009: 2nd Int. Conf. on Educ. Data Mining*. Cordoba, Spain, pp. 171–179.

Šimko, M. 2011. Automated Domain Model Creation for Adaptive Social Learning System. In *Inf. Sciences and Tech. Bull. of the ACM Slovakia*, Special Section on Student Research in IIT, Vol. 3, No. 2, pp. 119–121.

Šimko, M., Barla M., Mihál, V., Unčík, M., Bieliková, M. 2011. Supporting Collaborative Web-Based Education via Annotations. In *Proc. of ED-MEDIA: World Conf. on Educ. Multi-Hypermedia & Telecom.*, AACE, pp.2576-2585.

Suchal, J., Návrat, P. 2010. Full text search engine as scalable k-nearest neighbor recommendation system. In *AI 2010, IFIP AICT 331*, Springer, 2010, pp.165-173.

Tvarožek, M. 2011. Exploratory search in the adaptive social semantic web. In *Information Sciences and Technologies Bulletin of the ACM Slovakia*, Vol. 3, No. 1, pp. 42–51.

Uschold, M., Gruninger, M. 2004. Ontologies and semantics for seamless connectivity. *ACM SIG-MOD*. Vol. 33, No. 4, pp. 58–64.

Wong, W., Liu, W., Bennamoun, M., 2011. Ontology learning from text: A look back and in the future. *ACM Computing Surveys*. In press.

Wu, X., Zhang, L., Yu, Y. 2006. Exploring social annotations for the semantic web. In *Proc. of the 15th Int. Conf. on World Wide Web*, WWW'06. ACM, pp. 417–426.