Context-Based Retrieval in Digital Libraries: Approach and Technological Framework ♣

© Kurt Sandkuhl[1], Alexander Smirnov[2], Vladimir Mazalov[3], Vladimir Vdovitsyn[3],
Vladimir Tarasov[1], Andrew Krizhanovsky[2], Feiyu Lin[1], Evgeny Ivashko[3]

[1] School of Engineering at Jönköping University,
{kurt.sandkuhl, vladimir.tarasov, feiyu.lin}@ jth.hj.se
[2] St.Petersburg Institute for Informatics and Automation of the RAS (SPIIRAS)
{smir, aka}@iias.spb.su
[3] Institute of Applied Mathematical Research of the KarRC RAS (IAMR)
{vmazalov, vdov, ivashko}@krc.karelia.ru

Abstract

Digital libraries face similar challenges as enterprise information sources and the Internet: a fast growing amount of digital content requires enhanced ways of supporting information seeking. This paper presents an approach to context-based retrieval in Digital Libraries (DLs). The proposed approach includes creation of a profile representing general information demand of a user (abstract context), and use of ontology matching to identify the documents relevant to the operational context representing the current information demand of the user. A profile represents the user's interests as a DL reader and after creation is dynamically updated based on the changes in the user's interests. The identification of documents relevance is carried out by matching the user profile ontology against the digital library ontology. Semantic distance calculation is based on the use of a thesaurus.

1 Introduction

This paper aims at contributing to an improved relevance of results retrieved from digital libraries by proposing a conceptual framework for context-based retrieval. Digital libraries face similar challenges as enterprise information sources and the Internet: a fast growing amount of digital content requires enhanced ways of supporting information seeking. Capturing and exploiting preferences and other information about a user's information demand have been proposed as one contribution addressing this challenge. The use of context information has been found promising for this purpose.

One of the goals of context-based retrieval in DLs is to assess the relevance of documents for user needs [1, 3]. Nowadays, user faces problems of management and sharing of huge amount of documents saved in the DLs. The work presented in this paper proposes methodology and technological framework allowing the user to be provided with a set of relevant documents based on context-based retrieval. The paper concentrates on formalizing information demand of the user by profiling and matching the profile against an ontology describing documents in the digital library. The purpose to be achieved in our approach is an access of the user to the documents that are considered to be relevant for him/her in a particular situation (context).

Prerequisites for the approach are an ontological model describing typical interest of a DL user and a set of available DLs as document sources. The approach proposes a methodology assuming three stages. The first stage aims at creation of a context representing the user's information demand. The context is dynamically updated during the second stage. The third stage focuses on identification of documents relevant to the context (user needs). At this stage, the user profile ontology is matched against the digital library ontology. During matching, semantic similarity between the context and the shared ontology fragments is determined. Metrics of semantic similarity, used for comparison of semantically related words, similar ontologies, etc., are addressed in a number of algorithms, like HITS algorithm [6] for searching Internet pages using a structure of hyperlinks, PageRank algorithm [2], etc.

The paper is structured as follows. The introduction is followed by earlier work on understanding and capturing information demand with context models. A brief description of the overall framework is given in chapter 3. The next section describes the procedure of identification of relevant documents based on formalized context. The last chapter presents concluding remarks.

Proceedings of the 11th All-Russian Research Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» - RCDL'2009, Petrozavodsk, Russia, 2009.

2 Information Demand

The background for the proposed framework for context-based retrieval is earlier work on understanding and capturing information demand with context models. The notion of information demand is closely related to work in two areas: information logistics and information retrieval. Information retrieval aims at retrieving relevant information meeting the needs of a user, which are expressed by a query. In this context the aspect of relevance of information is of high importance. Saracevic [11] considers several types of relevance, e.g. algorithmic, topical and cognitive relevance. The underlying concept for algorithmic relevance is the relation between the query features and the search result. Topical relevance is the relation between aboutness of content objects and query. These two relevance concepts are important for retrieving information meeting the demand of a user, but do not contribute to explaining information demand as a concept. Here, we consider cognitive relevance of higher importance, which is the association between perceived information need of the user and information presented to the user based on retrieval results.

The main objective of the research field information logistics is improved information provision and information flow [7]. This is based on information demands with respect to the content, the time of delivery, the location, the presentation and the quality of information. The research field information logistics explores, develops, and implements concepts, methods, technologies, and solutions for the above mentioned purpose. A core subject of information logistics is how to capture the needs and preferences of a user in order to get a fairly complete picture of the demand in question. Principal approaches for this purpose are user profiles, situation-based and context-based demand models.

User profiles have been subject to research in information systems and computer science since more than 25 years. User profiles are usually created for functionality provided by a specific application. They are based on a predefined structured set of personalization attributes and assigned default values at creation time. Adaptation of such profiles requires an explicit adjustment of the preference values by the user. A situation-based approach was proposed for implementing demand-oriented message supply. The basic idea is to divide the daily schedule of a person into situations and to determine the optimal situation for transferring a specific message based on the information value. This approach defines a situation as an activity in a specific time interval including topics and location relevant for the activity. Information value is a relation between a message and a situation, which is based on relevance of the topics of a message for the situation, utility of the message in specific situations and acceptance by the user. Details and examples from collaborative engineering are given in [9, 10].

A context-based approach was proposed for use in enterprises or networked organizations. The basic idea is that information demand of a person in an enterprise to a large extent depends on the work processes this person is involved in, on the co-workers or superiors of this person and on the products, services or machines the person is responsible for. This led to the proposal to capture the context of information demand [8], i.e. a formalized representation of the setting in which information demand exists, including the organizational role of the person under consideration, work activities, resources and informal information exchange channels available.

3 Context-based retrieval in DLs

The framework of the context-driven retrieval is based on the use of an ontology-based model of the digital library (DL) and user profile representing typical information demand of the user. The documents to be found in DL are described through a user request expressing the current information need [14]. The retrieval request is modeled by two types of contexts: abstract and operational. *Abstract context* is an ontology-based model integrating knowledge on a general DL user and information about typical preferences of a particular user. *Operational context* is concretization of the abstract context based on the current need provided by the user's information request. Operational context is used by an ontology matcher for matching it against the ontology representing DL resources to find relevant documents (Figure 1).

Knowledge to be integrated in the abstract context is stored in an ontology describing the model of a general DL user that is combined with preferences of a particular user. The preferences are based on the initial information provided by the user to create a profile and information obtained dynamically by tracing the user's activities. The latter is used to keep the profile up-to-date and to assign weights to user preferences. The request data are used to identify the user profile fragment that corresponds to the current information need. The resulting operational context is an ontology fragment (or slice). A DL collection is represented with three types of ontologies: a document ontology, DL ontology, and shared ontology. A document ontology represents content of an individual document; a DL ontology formalizes content of all the documents stored in the DL; the shared ontology integrates the ontologies for all the libraries in the DL collection.

The ontologies can be created either manually by experts or in a semi-automatic way (e.g., [5]). According to the selected formalism, ontology A is defined as:

A = <O, Q, D, C>,

where:

O – a set of object classes ("classes")

Q – a set of class attributes ("attributes");

D – a set of attribute domains ("domains");

C – a set of constraints used to model relationships occurring in ontology representation formats / languages.

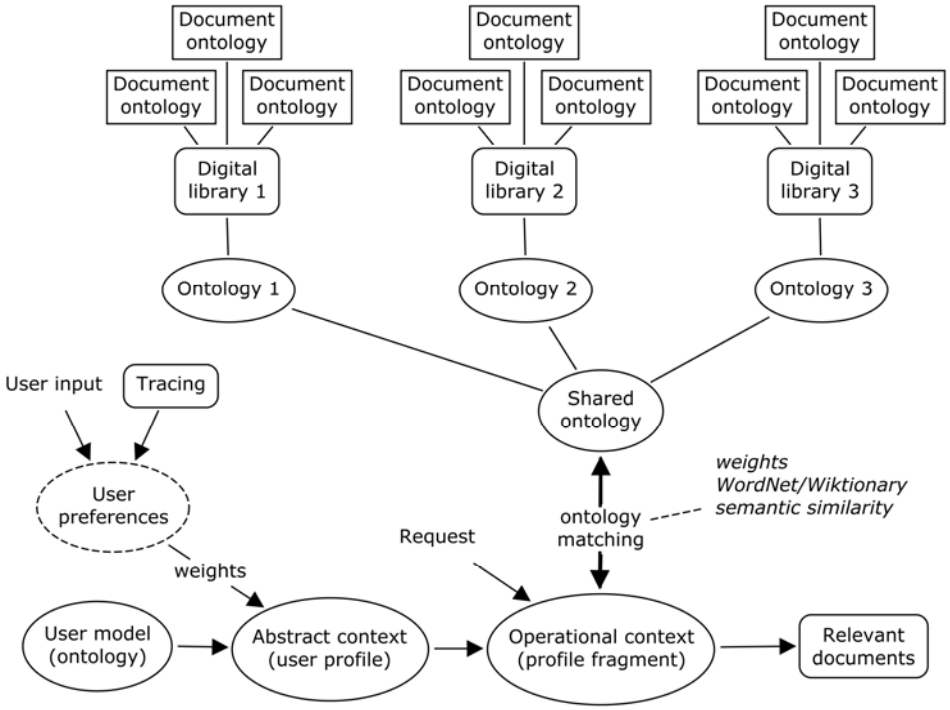

Figure 1. Conceptual framework of context-driven retrieval in DLs

For the DL user, the abstract context is a representational model of the user in a general view and the operational context is a current request description. For providing the DL users with documents relevant to the current request, a methodology and technological framework for context-driven retrieval in DLs have been developed. These methodology and technological framework have been developed within the conceptual framework of the context-driven information integration.

When abstract and operational contexts have been produced (i.e. the current situation has been described), the DL documents relevant to the current request are identified through ontology matching. WordNet and Wiktionary are used for improvement of semantic similarity algorithms during ontology matching. The resulting set of found documents with weights is presented to the user.

4 Context-based identification of relevant documents

In order to organize efficient identification of documents relevant to the current user request, the approach to context-sensitive access to information sources includes three stages (Figure 1): 1) creation of a profile representing general information demand of a user (abstract context); 2) updating the profile to adequately represent the current information demand of the user; and 3) use of ontology matching to identify the documents relevant to the operational context representing the current information demand of the user. A user profile is first created by ontological modeling of a DL user and then updated by behavioral modeling.

4.1 Constructing a profile of a DL user

Construction of a profile for a DL user draws upon our earlier work in competence modeling [15]. A competence model formalizes a person's skills and abilities, which are important for a certain task or situation. Competence models can be represented with ontologies. In a situation of document retrieval, the focus is on representation of the user's interests as a DL reader. These "reader's interests" can be described through professional interests and/or work role of the person in an organization.

As an example we can consider a user of a DL, which consists of scientific resources and aims at supporting workers in a research-oriented organization. A typical researcher would work in a research institute or university and would like to find documents relevant to their research interests. Hence, the main task is to represent research interests of the person. To do this, a user profile can be built based on papers published by the researcher and projects the researcher participated in. Each research paper/project can in turn be characterized by research fields relevant to the content of the paper/project. Additional research fields can be added by directly listing major research interest of the person and specifying fields connected to the scientific degrees or position.

To formalize research fields, different scientific taxonomies can be reused like the 1998 ACM Computing Classification System [16] or the Semantic Web Topic Hierarchy [12] in case of computer science.

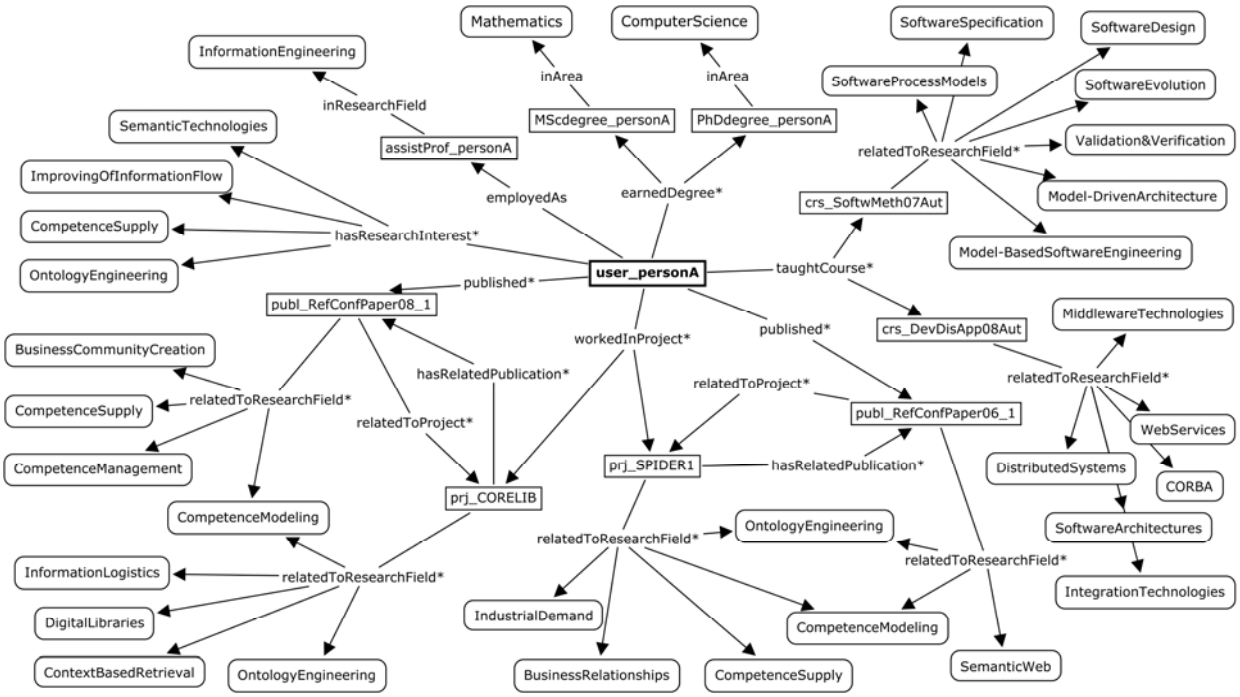

Figure 2. Example of a user profile for person A

If the person is engaged in teaching, then it is reasonable to include teaching topics in the model too. They can be included by connecting each taught course to the relevant research fields.

An example of a user profile is shown in figure 2. The profile includes research fields connected to two research papers, two projects, and two courses. Such a profile can be built manually or semi-automatically based on the general user information as well as the list of papers/projects for this person. Upon receiving a request from the user, the request can be mapped against the profile to identify the part (fragment) of the profile that corresponds to the user request.

4.2 Updating the user profile with behavioral modeling

After creation of a user profile describing topics of professional interest, the next task is to dynamically update the profile based on the changes in the user's interests. To trace these changes, a behavioral model is created representing current documents retrieved by the user. The behavioral model is constantly changed according to the user's activities. When there are many enough changes in the behavioral model, the user profile is updated to reflect the changes in the user's interests.

A behavioral model is built with the help of Markov chains [4]. The process includes tracing user searches to build a graph, constructing a classifier, and setting parameters. Using the profile shown in Figure 2, let us consider a possible behavioral model with respect to the project prj_CORELIB. Figure 3 depicts a simple example of a behavioral model, where each node represents transition from one topic of interest to another one. Every node is marked by the number of documents matching the appropriate attribute domain. Each topic can also be weighted based on the usage: access frequency, access date, etc. The transitions are extracted from a number of retrievals made by the user.

The example shows the difference between the original user's profile (see Figure 2, the part corresponding to prj_CORELIB) and the behavioral model (see Figure 3). The behavioral model related to prj_CORELIB contains two additional domains: CompetenceManagement and IndustrialDemand. But the domain IndustrialDemand appears only two times (this could show a weak interest that is it is not useful for the user in the context of prj_CORELIB) and CompetenceManagement – 6 times. This leads to the necessity to update the user's profile by adding the CompetenceManagement domain to the relation relatedToResearchField* (see Figure 4).

Thus, the creation of a behavioral model allows updating the user's profile according to the changes in the user's interests and therefore satisfying the user's current information demand.

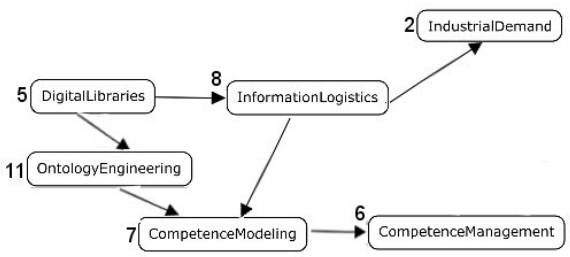

Figure 3. Example of a behavioral model of the user

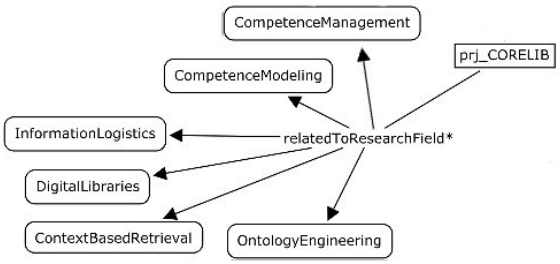

Figure 4. The updated part of person A's profile

4.3 Ontology matching based on WordNet and Wiktionary

Identification of DL documents relevant to the abstract and operational contexts is carried out at the stage of ontology matching. Since the current request is represented through the operational context, the aim is to sort documents, which are considered to be corresponding to the context, by relevance. Determining relevance is supposed to be based on measurement of similarity between the context and the shared ontology fragments, elements of which describe DL documents.

Ontology matching. A user profile is an ontology representing user preferences in terms of professional topics of interest and documents recently accessed. A digital library ontology describes the topics of documents stored in the library and relations between these topics. Hence, every document has one or several keywords or categories (like categories of wiki pages), which connect the document to the digital library ontology. After matching the user profile ontology against the digital library ontology, it is possible to predict potentially useful documents, which belong to the area of the user's interests.

This process consists of three steps:

1. Matching between the user profile ontology and digital library ontology. The result of this step is list *A* of entities of the digital library ontology corresponding to the user profile ontology.
2. "Closure" algorithm to find list *B*, which conforms to these three conditions:
    a. Entities of list *B* belong to the digital library ontology.
    b. Entities of list *B* do not belong to list *A*.
    c. Entities of list *B* are the closest elements of all the elements of the ontology to the entities of list *A*.
3. Enumerating a list of documents of interest for the user, which correspond to the entities of list B.

A set of ontology matching algorithms is based on thesauri (e.g., WordNet [17]). The comparison of different algorithms based on WordNet can be found in [18].

Semantic similarity. There are a lot of algorithms for semantic similarity, which are used for ontology matching. There is the following classification of ontology matching algorithms: internal and external [13]. An internal ontology matching algorithm exploits information that comes only with the source ontologies. An external ontology matching algorithm exploits external resources such as a domain ontology, corpus, thesaurus (e.g., WordNet, Wiktionary).

The Russian Wiktionary (the dump of the database as of January 2009) was parsed and the results were stored in a relational database (MySQL). Hence, the database of the parsed Wiktionary is the source data in the experiment [18].

The database of the parsed Russian Wiktionary has a better coverage than WordNet (247,580 words against 150,000). At the same time, WordNet consists of over 115,000 synsets while the total number of semantic relations in the database of the parsed Wiktionary is about 67,000 at this moment.

The experiment in [18] shows that the proposed method (Figure 5) is, in principle, capable of calculating a semantic distance between a pair of words in any language presented in Wiktionary (more than 200 in Russian Wiktionary). The comparison semantic distance between ontologies based on WordNet and Wiktionary raises an interesting question: whether the joint usage of Wiktionary and WordNet can improve calculation of the relatedness measure. This comparison is presented in [18].

5. Conclusions

The paper presents an approach to context-sensitive access to DL to help to identify documents relevant to a context (current situation). Capturing and exploiting preferences about a user's information demand have been proposed as one contribution. The approach includes three stages: (i) creation of a profile representing general information demand of a user (abstract context), (ii) dynamically updating the profile with behavioral modeling, and (iii) use of ontology matching to identify the documents relevant to the operational context representing the current information demand of the user.

The purpose of profiling (first stage) is to create a user profile by ontological modeling of a DL user. A profile represents the user's interests as a DL reader such as topics of professional interest and/or work role of the person in an organization. After creation the profile is dynamically updated based on the changes in the user's interests during the second stage. To trace these changes, a behavioral model is created representing current documents retrieved by the user.

The third stage focuses on identification of documents relevant to the current request (information demand of the user). The identification is carried out by matching the user profile ontology against the digital library ontology. The documents, which belong to the area of the user's interests, are sorted by relevance. Determining relevance is based on measurement of similarity between the context and the shared ontology fragments, elements of which describe DL documents.

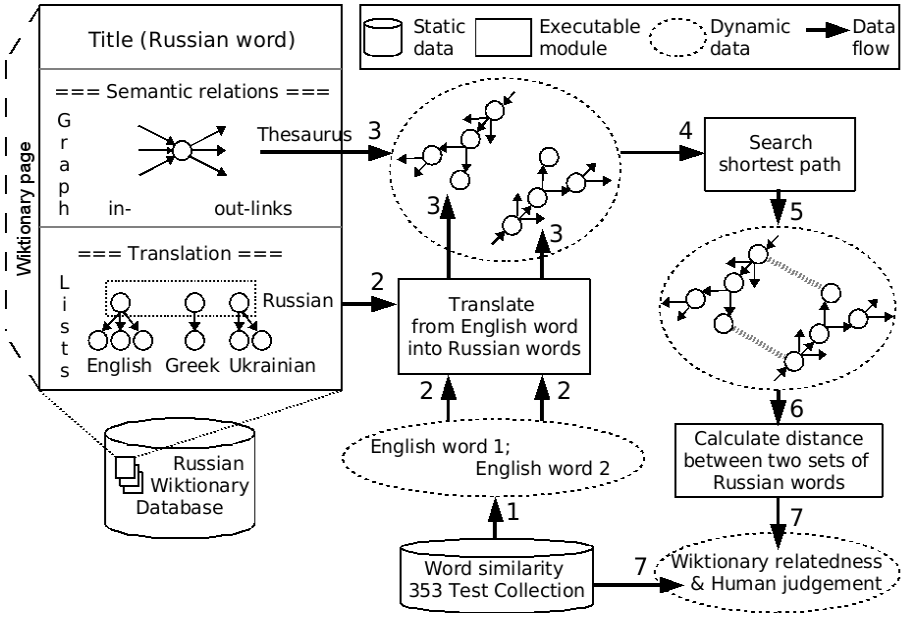

Fig. 5. Scheme of the experiment for calculating the semantic relatedness measure based on the Russian Wiktionary data

The shared ontology integrates the ontologies for all the libraries in the DL collection.

Our future work will focus on experiments with the proposed approach in a real setting. After initial modeling of several user profiles, experiments will be conducted in matching the digital documents against the user profiles by suing the semantic relatedness measure. The initial work in this direction has been started [18]. Further work with behavioural modelling to dynamically update user profiles is also an interesting direction that can enhance context-sensitive access to documents in DLs.

References

[1] B. Aleman-Meza, P. Burns, M. Eavenson, D. Palaniswami, and A. P. Sheth, "An Ontological Approach to the Document Access Problem of Insider Threat", *IEEE International Conference on Intelligence and Security Informatics (ISI-2005)*, Atlanta, Georgia, USA, 2005.

[2] S. Brin, and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", 1998. URL: http://www-db.stanford.edu/~backrub/google.html

[3] M. Ehrig, and A. Maedche "Ontology-focused crawling of Web documents", *Proc. of the 2003 ACM symposium on Applied computing,* Melbourne, Florida, 2003

[4] S. Jha, K. Tan, R.A. Maxion. Markov Chains, Classifiers and Intrusion Detection. In the proceedings of Computer Security Foundations Workshop (CSFW), June 2001.

[5] V. Kashyap, C. Ramakrishnan, C. Thomas, and A. Sheth, "TaxaMiner: An Experimental Framework for Automated Taxonomy Bootstrapping", *International Journal of Web and Grid Services, Special Issue on Semantic Web and Mining Reasoning*, Inderscience, 1 (2), 2005, pp. 240-266.

[6] J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, Association for Computing Machinery (ACM), 46 (5), 1999, pp. 604-632. URL: http://www.cs.cornell.edu/home/kleinber

[7] Kurt Sandkuhl: Information Logistics in Networked Organizations: Selected Concepts and Applications. Enterprise Information Systems, 9th International Conference, ICEIS 2008. LNBIP, Springer.

[8] Lundqvist, M. (2005). Context as a Key Concept in Information Demand Analysis. In Proceedings of the Doctoral Consortium associated with the 5th Intl. and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-05), 63-73. Paris, France.

[9] Meissen U., Pfennigschmidt, S., Voisard, A., and Wahnfried, T. (2004). Context- and situation-awareness in information logistics. In Postproceedings of Workshops of the International Conference on Extending Database Technology (EDBT), LNCS, Berlin/Heidelberg, Springer.

[10] Meissen, U., Pfennigschmidt, S., Sandkuhl K., Wahnfried, T. (2004a) Situation-based Message Rating in Information Logistics and its Applicability in Collaboration Scenarios. In Euromicro 2004 Special Session on "Advances in Web Computing", August 31- September 3, IEEE Computer Society Press.

[11] Saracevic, T.: Relevance Reconsidered r96. In Ingwersen, P.; Pors, N. O. (eds.): Information Science: Integration in Perspective. Royal School

of Library and Information Science, Copenhagen, Denmark, pp. 201-218, 1996.
[12] Semantic Web Topic Hierarchy, 2008. http://semanticweb.org/wiki/Semantic_Web_Topic_Hierarchy
[13] P. Shvaiko and J. Euzenat, 'A survey of schema-based matching approaches', Journal on Data Semantics, (IV), 146–171, (2005).
[14] A. Singhal "Modern Information Retrieval: A Brief Overview", *In IEEE Data Engineering Bulletin*, 24 (4), 2001, pp. 35-43. URL: http://singhal.info/publications.html
[15] Tarasov, V., Lundqvist, M. (2007). Modelling Collaborative Design Competence with Ontologies. In International Journal of e-Collaboration (IJeC): Special Issue on the State of the Art and Future Challenges on Collaborative Design, ISSN 1548-3673. IGI Publishing: Hershey, USA. Pages 46-62.
[16] The 1998 ACM Computing Classification System, 2009. http://www.acm.org/about/class/1998
[17] WordNet, 2006) "WordNet", 2006. URL: http://wordnet.princeton.edu
[18] Krizhanovsky, A. and Lin, F., 'Exploiting WordNet / Wiktionary in Ontology Matching', submitted to RCDL2009, Petrozavodsk, Russia.

Технология поиска в электронных библиотеках, основанная на контексте

К. Сенкюль, А.В. Смирнов, В.В. Мазалов,
В. Т. Вдовицын, В.В. Тарасов,
А.А. Крижановский, Ф.Лин, Е.Е. Ивашко

Электронные библиотеки в настоящее время сталкиваются с теми же проблемами, что и информационные системы предприятий, а также Интернет: быстро растущее количество электронных документов требует более совершенных методов поиска. В данной статье представлена технология поиска в электронных библиотеках (ЭБ), основанная на контексте. Предложенный подход предусматривает создание профиля, представляющего общие информационные потребности пользователя ЭБ (абстрактный контекст), и применения сопоставления на основе онтологии для распознавания документов, соответствующих операционному контексту, представляющему текущие информационные потребности пользователя ЭБ. Профиль представляет интересы пользователя как читателя ЭБ и после создания динамически обновляется на основе изменений интересов пользователя. Определение степени соответствия документов контексту выполняется через сопоставление онтологии профиля пользователя и онтологии ЭБ. Вычисление семантического расстояния основано на использовании тезауруса.

♣ The paper is based on research carried out as a part of several projects: project CoReLib supported by the Swedish Institute by grant # 01215-2007 and projects funded by grants of the Russian Foundation for Basic Research.