AN APPROACH TO PERSONALIZED SEARCH WITHIN DISTRIBUTED REPOSITORY OF VISUAL DATA

Andrzej GLOWACZ[1], *Tomasz Marcin ORZECHOWSKI*[2], *Andrzej DZIECH*[1]

[1]Department of Telecommunications, Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, Al. Mickiewicza 30, 30 059 Krakow, Poland
[2]Academic Computer Centre CYFRONET, AGH University of Science and Technology, Al. Mickiewicza 30, 30 059 Krakow, Poland

aglowacz@agh.edu.pl, tomeko@agh.edu.pl, dziech@kt.agh.edu.pl

Abstract. *In this paper, we introduce functional assumptions of the distributed repository of visual data that are one of key aspects of SYNAT framework for digital libraries. The proposed system will provide innovative tools and extend capabilities of current repositories. Within this task, the personalized search system is discussed. Especially three different approaches were proposed for this system, such as: CF, CBF and DF. Idea behind personalization is that there is currently lack of such functionality whereas end users would benefit from it.*

Keywords

Digital libraries, distributed repository, OAI-PMH, personalized search.

1. Introduction

The main goal of SYNAT project [1] is the establishment of the universal, open, hosting and communication, repository platform for network resources of knowledge to be used by science, education and open knowledge society. Developed system is intended to become widest publicly available repository service in Poland interlinking various digital libraries available and new data sets as well. Domains of use include but are not limited to: science, technics, education, history and culture. Project consortium consists of 15 institutions; among them top polish universities and national libraries.

One of key project objectives is the development of distributed repository of visual data that will provide innovative functions to the system. This task is conducted within four research areas. First one is related to development of distributed metadata repositories and methods of advanced search in the federation of repositories of visual data. Access interfaces will be based on common interfaces using SQI and OAI. Interlinking various databases furthermore requires unification of metadata description of stored multimedia objects. In addition, management of user profiles will enable complex search functionalities.

Second research goal is the creation of open database system. It will take advantage of ontology for design of database structure and allow for convenient integration with content-based multimedia indexing tools. Innovative elements include detection and prevention of distributed attacks against database system.

Another two objectives are strictly related to multimedia content and comprised of digital watermarking technologies and methods of high-quality content distribution. Considered digital watermarking is a set of suitable techniques that operate directly in multimedia content. Thus, useful functions related to data integrity and metadata verification; identification of distribution channel; or protection of sensitive data can be applied in repository. New watermarking techniques are also under research in this task. On the other hand, the content distribution system is focused on providing to end user the multimedia data in required quality. It will realize functions of adaptation and transcoding of visual data, and fragmentary object delivery (e.g. delivery of part of scanned manuscript or audio stream from video recording). Metadata system will store additional information about content quality.

Functional architecture of the repository and modules are presented in Fig. 1.

The remainder of the paper is organized as follows. Section 2 provides overview of the current distribution of digital content and existing systems for management of digital content. Section 3 introduces personalized search functionality within digital repository. Assumptions for information filtering, unified metadata, architecture and communications are presented. Section 4 concludes the work.

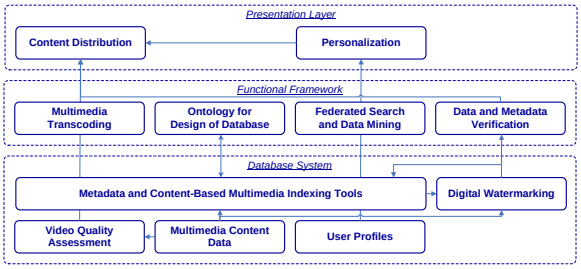

Fig. 1: Functional architecture of the distributed repository of visual data.

The remainder of the paper is organized as follows. Section 2 provides overview of the current distribution of digital content and existing systems for management of digital content. Section 3 introduces personalized search functionality within digital repository. Assumptions for information filtering, unified metadata, architecture and communications are presented. Section 4 concludes the work.

2. Related Work

Although there are many search systems for digital libraries, the personalized search approach was not so far taken into account. This was a result of a simple and correct assumption that if the current catalogue systems were so far verified during years, the digitization could be the only a transfer of the existing model into the on-line system.

The development of IT technology leads to increasing growth of interest in electronic materials. It is worth quoting a following fact: according to data revealed by the Amazon, as of 1 April 2011, a shop sells an average of 105 e-books to 100 paper books [2]. The calculations take into account also the paper books that have no electronic equivalent, and it was not taken into account the distribution of free e-books. This fact also surprised the president of Amazon. Jeff Bezos said, that he hoped that such a state would happen some when, but no one expected that it had happened this time. Especially, that books are sold for 15 years, whilst e-books only from less than 4 years [2], [3]. Although the data refers to the U.S. Amazon, but e-books are also popular in the European branches of the store. Gordon Willoughby, the European director of Kindle, said, that e-book sales by the British Amazon.co.uk were unexpected. For example, in the UK, where e-books have been sold for only 9 months and hard-books for 13 years, now they sell 242 e-books for every 100 hardbacks [2], [3].

Two different steps in the process of digitization of typical libraries could be distinguished:

- digitalization of librarian catalogues,
- digitization of documents and other librarian resources (paintings, films, etc.).

Digitalization of resources stored in libraries allows for implementation of additional methods for indexing content and thus it for generation of the appropriate indexes regardless of previously used methods of cataloguing. In other words, the metadata representing librarian resources can now be enriched with the results of the process of libraries resources indexing.

At the same time, the increased interest in distribution of video content can be observed. Although the terrestrial and satellite digital TV providers have offered VoD services for last years, only now, thanks to the increase of bandwidth and the number of broadband Internet users, the implementation of such services is observed. The description of the most popular systems for management of digital libraries is summarized in Tab. 1. The technology revolution in mobile telephony both in terms of transmission speed and equipment enables current smartphones' users not only to make video calls, but also soon to watch TV in the DVB-H standard. Existing technologies of image and sound analysis allow for indexing of the video and audio sequences. They are used to describe the content, and even the character of the given sequence. Regarding to the description of the nature of content, it is worth to mention, that the web portals, which offer the music tailored to the listeners' mood, exist nowadays.

Personalization was not a key element in any systems managing digital libraries that were shown in Tab. 1, so the proposal presented in this paper has an innovative character. Despite significant differences in the characteristics between distributed e-learning systems, and digital libraries, adaptation of such a solution to the requirements of repositories seems to be desirable and will offer a real personalized search for digital libraries' users.

Tab.1: Summary of selected systems used for management of digital content.

System's name	Description	Supported technologies	License type
CONTENTdm	The system for management of digital collections. The software can handle the storage of different digital object including: local history archives, newspapers, books, maps, slide libraries or audio/video. It supports also Optical Character Recognition (OCR) on text documents to enable full text searching [4].	Unicode, Z39.50, Qualified Dublin Core, VRA, XML, JPEG2000, OAI-PMH.	Commercial software

dLibra	The system for building digital libraries, developed by the Poznan Supercomputing Centre - Network (PSNC). The intention was to implement the framework for digital libraries in PIONIER network and to established the Polish Federation of Digital Libraries. This system is well known and commonly used in Poland [7].	MARC, Dublin Core, OAI-PMH	Commercial software
DSpace	The package provides the tools for management of digital assets and it enables open sharing of content that spans organizations. It supports a wide variety of data, including books, theses, 3D digital scans of objects, photographs, film, video, research data sets and other forms of content [5].	METS, OAI-PMH	Open source software under BSD License
ExLibris DigiTool	The software is used for managing and showcasing of Digital Collections and Institutional Repositories. It allows for: cataloguing, managing, sharing, searching, and retrieval of digital collections [6].	MARC 21, Z39.50, Qualified Dublin Core, METS, Z39.87-2002, OAI-PMH	Commercial software
FEDORA	Flexible Extensible Digital Object Repository Architecture that provides a management layer for digital objects based on content models, which represent data objects, or collections of data objects (e.g., digital images, XML files, metadata) [8].	REST/SOAP, SPARQL OAI-PMH	Open source software under Apache License 2.0
Greenstone Digital Library	It is a suite of software for building and distributing digital library collections. It enables the construction and presentation of information collections with effective full-text searching and metadata-based browsing facilities. The system is extensible: software "plugins" accommodate different document and metadata types [9].	Z39.50	Open source software under Apache License 2.0
INVENIO	It is a software suite enabling to run digital library or document repository on the web. The technology offered document ingestion through classification, indexing, and dissemination [10].	MARC21, OAI-PMH 2.0	Open source software under GPL License 2.0
SimpleDL	It is a digital collection management software that supports different type and format of handling documents, such as: images, videos, audio files [11].	OAI-PMH	Commercial software

3. Personalized Search

Personalized approach assumed the introduction of different methods to obtain, during the search, results tailored to the needs of specific end-user. There are different approaches in the implementation of personalized search, but generally the approaches based on filtration results could be distinguished. Filtration is to reject or change the total conformity value of the respective objects (from the result set) considering both conformity to the given query and given user's preference.

3.1. Information Filtering and Unified Metadata

It is intended to establish modularized framework of the system (Fig. 2) that could support all mentioned above approaches. The details of these well-known solutions are deeply analysed and described in many scientific articles, compare: [15], [16], [17]. The most frequently used filtering methods include [15]:

- CF (Collaborative Filtering),
- CBF (Content-Based Filtering),
- DF (Demographic Filtering).

It is intended to establish modularized framework of the system (Fig. 2) that could support all mentioned above approaches. The details of these well-known solutions are deeply analysed and described in many scientific articles, compare: [15], [16], [17]. There are two the most popular standard describing metadata of digital resources:

- describing e-learning objects: IEEE LOM [12],
- describing digital libraries' objects: MARC [13].

The BSS (Broker Service System), (Fig. 2) is responsible for parsing all these type of metadata standards. The details of components of the proposed framework are presented in the subsection: Architecture. Another metadata standard, that is intended to be supported by the ASMS (Advance Search Management System), (Fig. 2), is Attention Metadata [14]. The main purpose of this standard is to exchange the information describing users' attentions to search results.

3.2. Architecture

The framework of the proposed system is shown in Fig. 2. We can distinguish following parts of the framework:

- Search System – the main part that is responsible for gathering queries from end-users, transferring them to another module such as BSS or its extension: Advance Search System. Then it is also responsible for presenting the results to end-users.

- Broker Service System – is responsible for organising the data for query to be sent via the broker layer to repositories. The BSS supports different standards describing metadata (including: IEEE LOM and MARC 21), different query languages (including: CQL and S2QL), as well as different services (SOAP, REST) and interfaces: SQI, SRU and SRW.
- Broker – is responsible for connection to repositories and organize the federated search.

Advance Search System – is the core part of personalization. The positioning of found results is the core aspect of the search done by given end-user, because it can fulfil requirements of personalized search. This system consists of some sub-modules including: module responsible for storing end-user profiles, module that collects end-users behaviour data, module that collect the end-users' attention to found results. All these modules lead to offer the CF oriented personalization, where end-users are divided into clusters according to the similarity, either of their behaviour or of their profiles.

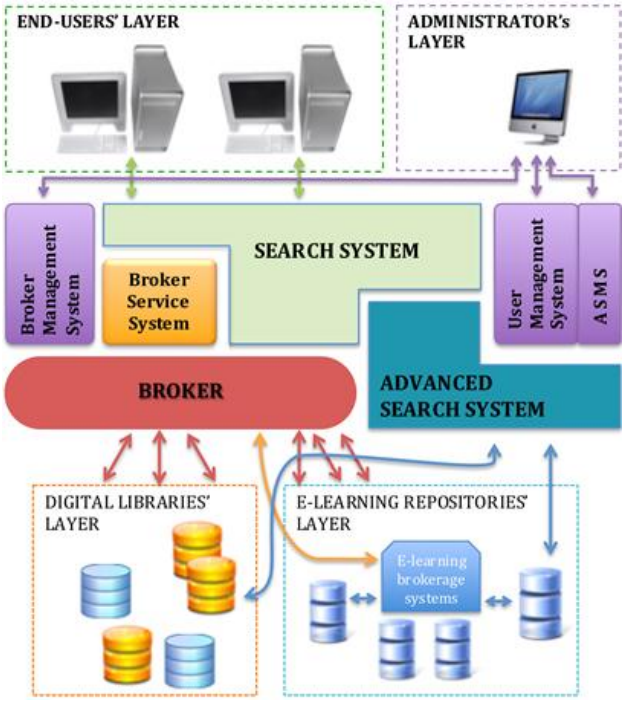

Fig. 2: The model of the framework of the personalized search system.

There are also some modules used by the administrator to manage the broker, users accounts and advance search systems (Fig. 2). They are called: Broker Management Systems, User Management System and Advance Search Management System.

3.3. Communications

The Communication done by broker (Fig. 2) includes: communication with repositories that support SOAP Services including implementation of SQI over SOAP (S2QL support) and the successors of Z39.50 standard (supporting CQL): SRU (Search/Retrieve via URL) and the SRW (Search/Retrieve Web service).

Moreover, many of repositories support OAI-PMH, so it is also planned to extend the module of Advanced Search by introduction following sub-modules: collecting module and indexing metadata module. The first one will be responsible for the OAI-PMH communication with repositories, data harvesting and storing, whilst the second one will be responsible for the metadata indexing (including both IEEE LOM and MARC21 standards) and offering the recommendation utilizing Content-Based Filtering.

4. Conclusion

The proposed solution consists on introduction the personalization search for digital libraries' users. The personalized search is not commonly offered by any of existing software for managing digital libraries. System will utilize the collaborative filtering by organize end-users into two different kinds of clusters on the base of two different profiles:

- Dynamic: created according to end-users' behavior,
- Static: connected to end-users explicitly defined profiles.

The proposed system will not only offer personalized search for digital libraries, but also for the content stored in e-learning repositories. It is also planned to offer personalization extension by implementing both Collaborative Filtering and Content-Based Filtering and finally obtain the best effect for Personalized Search.

Acknowledgements

Work financed by The National Centre for Research and Development (NCBiR) within SYNAT project no. SP/I/1/77065/10.

References

[1] SYNAT Project. *Synat: System nauki i techniki* [online]. 2012. Available at: http://www.synat.pl/.

[2] GABATT, Adam. Amazon and Waterstones report downloads eclipsing printed book sales: Success of Kindle electronic reader prompts rapid rise of ebooks, with UK enthusiasm outstripping US. In: *The Guardian* [online]. 2011. Available at: http://www.guardian.co.uk/books/2011/may/19/amazon-waterstones-ebook-sales.

[3] PRIGG, Mark. After only nine months of Kindle, ebooks outsell hardbacks at Amazon. In: *The Guardian* [online]. 2011. Available at: www.thisislondon.co.uk/standard/article-23952130-after-only-nine-months-of-kindle-ebooks-outsell-hardbacks-at-amazon.

[4] CONTENTdm: Digital Collection Management Software. *OCLC: The world's libraries* [online]. 2012. Available at: http://www.contentdm.org/.

[5] DSPACE. *DSpace* [online]. 2012 [cit. 2012-11-01]. Available at: http://www.dspace.org/.

[6] ExLibris: DigiTool. *ExLibris* [online]. 2012. Available at: www.exlibrisgroup.com/category/DigiToolOverview.

[7] DLibra: Digital Library Framework. *DLibra* [online]. 2012. Available at: dlibra.psnc.pl.

[8] FEDORA: Flexible Extensible Digital Object Repository Architecture. *Fedora Commons* [online]. 2012 [cit. 2012-11-01]. Available at: http://fedora-commons.org.

[9] Greenstone: Digital Library Software. *Greenstone* [online]. 2011. Available at: www.greenstone.org.

[10] Invenio. *Invenio* [online]. 2012. Available at: invenio-software.org.

[11] SimpleDL: Digital Libraries Simplified. *SimpleDL* [online]. 2012. Available at: simpledl.com.

[12] IEEE P1484.12.3. *Draft Standard for Learning Technology: Extensible Markup Language (XML) Schema Definition Language Binding for Learning Object Metadata*. New York: IEEE, 2005. Available at: ltsc.ieee.org/wg12/files/IEEE_1484_12_03_d8_submitted.pdf.

[13] Marc Standards. *Marc 21* [online]. 2012 [cit. 2012-11-01]. Available at: http://www.loc.gov/marc/.

[14] OCHOA, X. and E. DUVAL. Use of contextualized attention metadata for ranking and recommending learning objects. In: *Proceeding of the 1st International workshop on Contextualized Attention Metadata (CAMA'2006).* Arlington: ACM, 2006, pp. 9-16. ISBN 1-59593-524-X. DOI: 10.1145/1183604.1183608.

[15] MONTANER, M., B. LOPEZ and J. L. DE LA ROSA. A Taxonomy of Recommender Agents on the Internet. *Artificial Intelligence Review.* 2003, vol. 19, iss. 4, pp. 285-330. ISSN 1573-7462. DOI: 10.1023/A:1022850703159.

[16] ADOMAVICIUS, G. and A. TUZHILIN. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *The IEEE transactions on knowledge and data engineering*. 2005, vol. 17, iss. 6, pp. 734-749. ISSN 1041-4347. DOI: 10.1109/TKDE.2005.99.

[17] PAZZANI, M. A Framework for Collaborative, Content-Based, and Demographic Filtering. *Artificial Intelligence Review.* 1999, vol. 13, iss. 5-6, pp. 393–408. ISSN 1573-7462. DOI: 10.1023/A:1006544522159.

About Authors

Andrzej GLOWACZ received his Ph.D. in Telecommunications from the AGH University of Science and Technology in 2007. He is currently an Assistant Professor at the AGH University. Andrzej Glowacz has been working on numerous commercial projects, and EU research projects including INSIGMA (as Deputy Project Coordinator), INDECT, GAMA, OASIS Archive, CARMEN, DAIDALOS, DAIDALOS 2, EuroNGI, and EuroFGI. His main professional areas of interest are intelligent information systems, multimedia systems, pattern recognition, wireless QoS, modern transport protocols, and advanced systems programming. He is the author of over eighty scientific papers and technical reports; he also serves as a reviewer of international journals and conferences.

Tomasz Marcin ORZECHOWSKI received his M.Sc. in Computer Science in 1999 and Ph.D. in Telecommunications in 2010 from the AGH University of Science and Technology. He is currently an Assistant Professor at the AGH-UST. His research interests include: distributed programming, mobile agent technology, databases, recommender systems and information retrieval. He is the author of numerous publications. He has been a member of the IEEE Communications and Computer Societies for over than 10 years. He is the member of several conference committees.

Andrzej DZIECH graduated from the Leningrad Electrotechnical Institute at the Faculty of Automation and Computer Science, specializing in automation. in 1970 He received his Ph.D. in Telecommunications from the same institute in 1973. He received the title of professor in 1986. In 2002-2005 he was the Chairman of the Committee of the Section of Telecommunications Research in Poland. Currently, he is the expert of European Union, evaluating scientific projects in the UE 7th Framework Programme. He is also the main coordinator of EU FP7 INDECT Project, titled: "Intelligent Information System Supporting Observation, Searching and Detection for Security Citizens in Urban Environment". He also led EU project: CALIBRATE, FP6-IST - Strengthening the Integration of the ICT research effort in an Enlarged Europe.