Towards Bridging the Gap between Personalization and Information Extraction

Nirmala Pudota, Paolo Casoto, Antonina Dattolo, Paolo Omero, Carlo Tasso
Department of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206 - Loc. Rizzi
Udine, Italy
Email: {$nirmala.pudota, paolo.casoto, antonina.dattolo, paolo.omero, carlo.tasso$}@dimi.uniud.it

***Abstract*—In this paper we propose to integrate Information Extraction and Adaptive Personalization in order to empower information access and Web search experience. We describe the PIE (Personalized Information Extraction) architecture which exploits zz-structures for organizing information and user profiles for capturing personal user interests in digital libraries. We apply our model to Bibliomed system in order to extend its functionalities.**

I. Introduction

The explosive growth and popularity of the World Wide Web has resulted in a massive amount of information sources on the Internet, creating a scenario where the answers to information needs of the users are available online somewhere in some format; but in order to find the appropriate information users need to scan through endless list of digital data. Different typologies of users explore the Web in various ways according to their requirements and experiences; some users, for instance, may survey an area of knowledge to get a general understanding on it, while others to look for specific information. In either of the cases, they need to access and analyze all the documents available and this process is time consuming. For these reasons, they normally tend to compromise themselves with the information they have received. This clearly indicates the presence of information overload [1], [2]. Personalized web content [3], [4] is one of the proposed solutions to solve this problem.

Moreover another feature of information available on the Web makes difficult identify opportune, automatic and effective methodologies of access and retrieval: the most part of information is present in the form of unstructured free text, written in natural languages. Examples are blogs, forum, corporate memos, research reports, emails, blogs and historical documents [5].

According to recent studies more than 80% of queries submitted by users to search engines are estimated informational in nature [6]. This means that most of them could be answered properly by providing structured and normalized form of information, like to key notes of entities, price lists of items for sale, document summaries. The purpose of *Information extraction* (IE) is to structure the possible unstructured text; in other words, IE is the process of populating a template of structured information starting from unstructured or loosely formatted text, which can be given directly to user or can be stored in a database for further processing [5], [7].

Our research is devoted to IE applied to text processing, but it can be generalized to other kinds of multimedia data, like images, audio and video contents.

This paper is organized as follow: Section 2 illustrates Information Extraction and the related areas of Information Retrieval and Text Mining; Section 3 introduces User Modeling concepts focusing on some aspects of personalization, while Section 4 illustrates the proposed PIE architecture. Finally, Section 5 concludes the paper.

II. Information Extraction and Related Areas

A. Information Extraction

Digital text is abundant and every day thousands of new Web pages are created, loaded and made available in various forms. Advanced search functionalities, which allow the user for detailing search at a finer granularity, are effectively used in current search engines, bur several activities, like fielded searches, join-based structured queries, text mining and support to user decision, require more structured ways to represent the textual data [5]. An example of such activity is represented by a user interested in collecting and comparing the prices of a given set of goods, for instance a brand new mobile phone, proposed by a set of e-commerce sites. A table reporting the price of each seller may be more useful and easier to understand than a list of Web pages user needs to browse. IE is aimed at recognizing and extracting the information required for further processing from unstructured text, accordingly to a specific extraction template. This information can be stored in databases or spreadsheets for later analysis or can be given directly to the user.

The activities of IE and definition of extraction templates are strongly coupled with the application domain in which they are involved and require a great amount of domain knowledge for effective performance. Portability of IE methodology and artifacts between different domains is a critical issue. For example, an IE system designed for a financial domain might extract entities like company names, locations, dates and financial figures like currency amounts and percentages, while an IE system designed for a medical domain might extract

entities like names of diseases, virus, drugs, proteins and information about treatment from clinical records.
The Message Understanding Conferences or Competitions (MUCs) have inspired early works in IE [8], [9] in response to the opportunities offered by the enormous quantities of online texts.

B. Information Extraction and Information Retrieval

A natural association links IE with Information Retrieval (IR), but a basic difference distinguishes IR from IE [10]: IR aims at retrieving all and only the documents storing information *relevant* to the user's information needs, while IE aims at extracting text which matches a template; either it is manual or automatic way, the goal of IE is to search for words, paragraphs, or text snippets contains searched information matched to the specified template and present it in a more organized and structured form. This means that the central notion of IR is relevance, while that of IE is *information structure*. The former is represented through a query (or, more generally, by some implicit or explicit input from the user), whereas the latter is represented by a *template*. IE and IR can be used in a combined way: IE can use IR to select relevant documents for further analysis; from another point of view, the structure templates or normalized databases filled by IE can be utilized by IR for more flexible results. Also, IE might be useful as a preparatory step for IR as well as for post processing [10].

C. Role of Information Extraction in Text Mining

Information Extraction represents a starting point to analyze unstructured or natural language texts; specifically, by combining natural language processing tools and lexical resources, IE can provide effective mechanisms for mining documents of various domains [7].
In particular IE is aimed at extracting from given documents the occurrences of particular entities, relationships between them and events which are able to describe the evolution of both entities and relationships. IE provides the set of linguistic features required to organize relevant information, defined by a specific extraction template, in a structured way. Structured information can be further processed in order to identify latent or emerging concepts, such as relationships or events, which are not covered by the extraction template. Such approach is known with the term *Text Mining*, and can be seen as a specialized Data Mining activity focused on textual contents.

III. Personalization through user profiles

As the amount of information accessed by users became large, the identification of relevant pieces of information, according to users information needs, became a critical task. To overcome this information overload, several conceptual and application frameworks [1]–[4], [11] have been proposed in terms of personalized Web content and personalizing Internet information services. The term "personalization" is employed in a wide variety of contexts, and a general definition of the associated concept is given in [12] as the task aimed at:"*bringing the right piece of information to the right person in the right time*" . User models and personalization methods have been addressed in different fields, like IR systems and database systems (DB).
In query personalization, user preferences are stored as user profiles and dynamically used by integrating those profiles that are relevant to the query. Recently query personalization is generating huge interest in both IR and database research communities.

Personalized content retrieval is aimed at improving the process of retrieval by considering the user profile as an active part during the relevance evaluation process. This approach leads from a "one fits all" criterion, which is nowadays the way the most common search engines manage the complexity of retrieval to a user centered retrieval.
Nevertheless, as not all human preferences are similar and they usually change depends on situations. It is well known that user preferences are complicated, multiple, ever changing and they should be understood within a context which depends on user goals [11].
In this work, we are more interested to the aspects related to user profiling for personalization, which refers to the collection of information about user for generating user profiles. User profiles can be acquired in several different ways, by means of implicit and explicit interaction with users, for instance by asking them to fill-in a profile form (explicitly) or by automatically tracking their activities on the Web (implicitly) [2].

IV. Information Extraction and Personalization: Bridging the Gap

Over the past decade much progress has been made in the field of IE and significant attention was laid on it; this has lead to propose and develop many efficient and effective techniques to extract information for a given text corpus (e.g., a collection of news articles, Web pages, blogs, emails) [13]. As a result of these efficient and effective techniques, IE has seen a tremendous progress so much that millions of facts extracted from the Web are now turned into knowledge bases. But, it is not clear how effectively these knowledge bases are browsed and how they support common user information needs [14].

We claim that, integrating IE and adaptive personalization techniques and user profiles can greatly impact and improve Web information access and user's search experience.

We propose an innovative architecture aimed at personalizing the information extraction process; our approach is called *Personalized Information Extraction* (PIE). A general architecture implementing PIE is shown in Figure 1.
Our proposed architectural model can be specialized on several different application domains and content sources. In this paper we focus our attention in the medical domain, integrating our ideas with the features provided by BiblioMed[1], an application devoted to personalized access to an heterogeneous set

[1] http://bibliomed.bib.uniud.it/

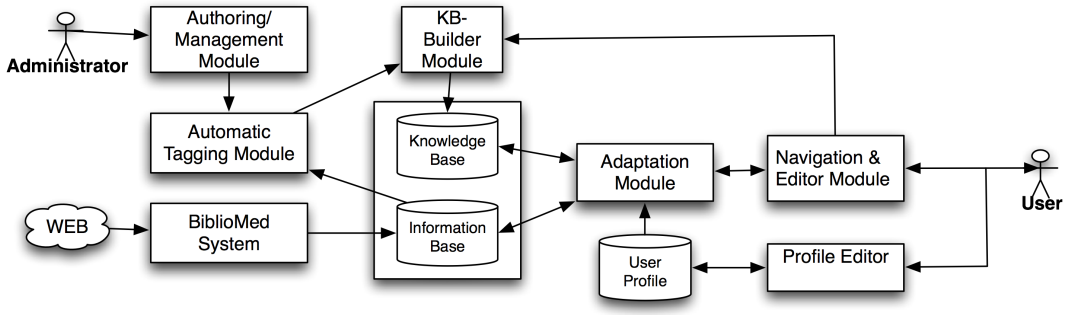

Fig. 1. The proposed PIE architecture.

of medical digital library [15].
The following modules constitute the proposed architecture:

1) The *Information Base* (IB) represents the set of textual documents which can be accessed and organized by the user. In our case study, the IB is constituted by the documents gathered by the BiblioMed meta-search-engine, a module of BiblioMed devoted to monitoring newly available contents from an heterogeneous collection of Web sources. In particular the information providers of Bibliomed are: medical databanks, such as PubMed-Medline, online medical journals, library catalogs, medical forums, blogs and directories.
2) The *Automatic Tagging Module* (ATM) is used to enrich the documents of the IB with a series of annotations (or tags) which are useful to arrange the documents into concept maps. The ATM provides large-scale metadata annotations as an important step towards realizing the Semantic Web; in particular, it performs two different kinds of tagging activities:
   a) *Information Extraction*. The goal of Information Extraction is aimed at extracting named entities in order to automatically perform tagging of textual documents. In the specific domain of our case study we are interested in extracting both domain independent (e.g. person names, organizations, locations, dates) and domain dependent (e.g. diseases, proteins, anatomical issues included in the Medical Subject Headings MeSH) entities. In addition we are interested in extracting meta-information about the retrieved documents, such as information about the authors, their affiliation or the specific publication where the document has been published.
   These tags are useful not only in building knowledge bases but also they help the end-user in knowledge organization and access the information units by building his/her own conceptual space. For the annotation and named entity extraction task we have used freely available information extraction system GATE[2] which offers pre-defined transducer for the recognition of different named entities, a recognizer of English verbal phrases, a Gazetteer (works on lists of names), and language processing tools like tokeniser, sentence splitter, POS tagger.
   b) *IFT-Automatic Tagging*. The IFT-Automatic Tagging algorithm [16], previously applied in the field of Information Filtering, is used in addition to previously defined approaches in order to tag the documents. The input document is processed and transformed into a semantic network, where cells, representing terms, are linked by weighted arcs, representing the semantic relation of co-occurrence between the terms at each edge of the arc. A ranking criterion is used to establish the n most relevant terms of a document. The IFT-Automatic Tagging algorithm is not aimed at identifying terms with a specific semantic characterization; the terms selection is based on statistical evaluation of the textual features. IFT will extract relevant terms from a text in an unsupervised way, without referring to any kind of resource base. Examples of tags for documents related to Cardiology are *cardiomyopathy, valves, cardioprotection, hypertrophy, hypertension, bypass, heart failure*. More details about the IFT algorithm can be found in [16].
3) *Knowledge Base*. The Knowledge Base (KB) is a repository of the knowledge, extracted by the ATM for each document of IB. The knowledge in KB is organized as a conceptual map whose implementation is based on a specific knowledge representation model. In order to attain this specific representation model, the KB-Builder module has been introduced in our architecture. This module implements the business logic which is needed to move from a set of annotated documents to a conceptual map representing the extracted knowledge. KB-Builder module structures and organizes information deriving from ATM by means of zz-structures [17], a

[2]http://www.gate.ac.uk/

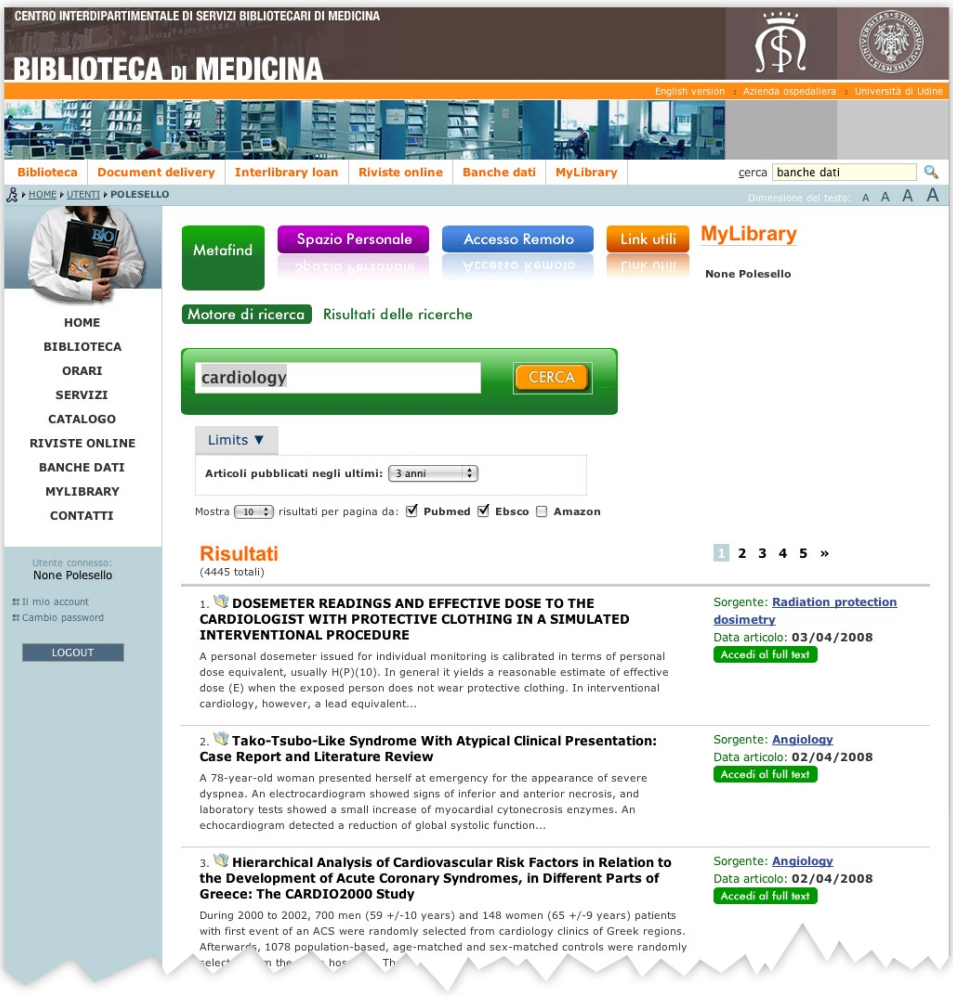

Fig. 2. Results from the BiblioMed Meta-Search Engine for the given search query.

graph-centric system of conventions for data and computing. A zz-structure can be thought of as a space filled with cells. Cells are connected together with links of the same color into linear sequences called dimensions. A single series of cells connected in the same dimension is called rank, i.e., a rank is in a particular dimension. Moreover, a dimension may contain many different ranks. For any dimension, the degree (no. of in/out links of a given color) of each cell cannot be greater than 2; this restriction ensures that all paths are non-branching, and thus it provides the simplest possible mechanism for traversing links. Zz-structures have been applied with success in different fields, as electronic music [18], data grid systems [19] and e-learning environments [20]. In our approach, knowledge representation cells are linked accordingly with the existence of a tag-relation between them. For instance, a cell corresponds to a synthetic description of a Web document and a link connects two cells whenever the same tag has been assigned to the two documents. The tag-relations can be defined by the ATM or by users, using the Navigation & Editor module. In such a way the document collection is augmented with dimensions derived from the ATM process to comprise in a possible new (navigation) paths. Figure 2 is shown the output of the BiblioMed system obtained by applying the search query “Cardiology”, and by setting the field “Articoli pubblicati negli ultimi:”(articles published in last) to “3 anni”(3 years). An example of derived zz-structure is shown in Figure 3. The cells $a_1, \ldots, a_{25}$ represent the first 25 results obtained in Bibliomed (see Figure 2); connections reflect similarities among articles in terms of $topic$, publication $year$, $author$ and name of the $journal$ the article was published.

Normal, thick, dotted and dashed lines represent respectively the dimensions D^{topic}, D^{year}, D^{author}, $D^{journal}$ generated from the set of annotations assigned by the ATM.

Each dimension can contain one or more ranks; for instance, in the example, D^{year} contains two ranks:R^{year}_{2008} containing the cells a_1, a_2, a_3, a_8, a_9, a_{10}, a_{11}, a_{21} and R^{year}_{2007} the cells a_4, a_5, a_6, a_7, a_{12}, a_{13}, a_{14}, a_{15}, a_{16}, a_{17}, a_{18}, a_{19}, a_{22}, a_{24}, a_{25}. In this way, depending on

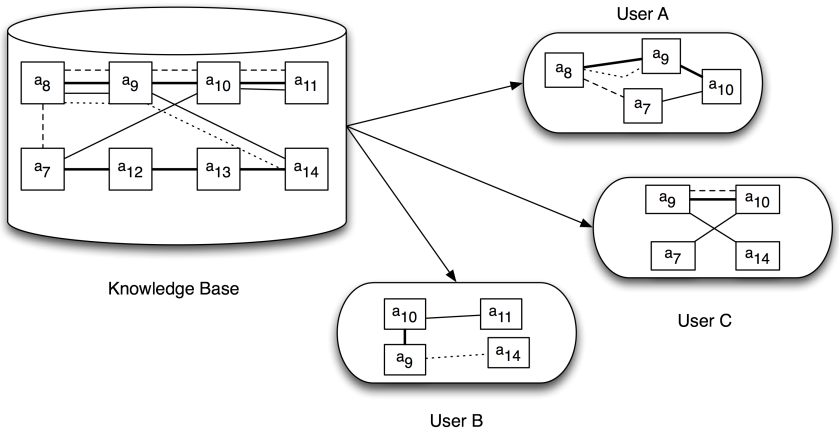

Fig. 4. Personalized sub-concept maps based on three different UPs

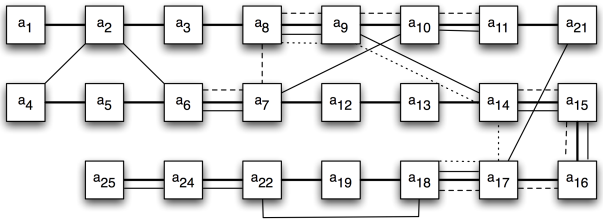

Fig. 3. An example of zz-structure

the specific tags, dimensions or ranks, it is possible to perform different abstractions, relevant to different user needs and perspectives. The union of zz-structure-based concept maps generate the user concept space: it can be defined [20] in terms of a multi-agent system constituted by five types of agent classes, respectively related to concept maps, dimensions, ranks, composite and atomic cells. These five agent classes represent five abstraction levels in the user concept space. Concept agents split the concept space into topic-related zz-structures; they know and directly manipulate dimensions and isolated cells, including concepts and relationships between concepts (organized in dimensions). Dimensions agents, uniquely identified by dimensions color, know and manipulate their connected components (ranks). Ranks know and coordinate the cells and the links that connect them; finally, composite cells agents contain concept maps related to more specific topics, while atomic cells agents are primary entities and directly refer to documents. Agents collaborate in order to manage, maintain and visualize concept spaces, and/or part of them.

4) The *User Profile* (UP) is assigned to each user in our model. Such profile is constituted by different kinds of information collected both implicitly and explicitly. Each user can access a wide amount of information units and selects cells and links he/she wants to visualize; such entity selection is stored into the UP. Every entity, both cells or arcs, of the zz-structure used to represent the KB is characterized by a specific user access permission. The set of permissions granted to a specific user by his/her UP defines the personalized user concept space, a subset of the overall KB.
5) The *Profile Editor* module is responsible for building and managing user profiles devoted to represent specific individual user interests. The contents of profiles can be directly provided by the user (explicit personalization) or it can be result of (automatic) analysis of user navigation history he/she makes by using Navigation & Editor Module; examples of user activities tracked by navigation history are: documents observed and time spent on them, selected dimensions and views, submitted queries, manually added documents and tags.
6) The *Adaptation Module* (AM) is responsible to extract information from the KB according to specific interests of the users which are explicitly/implicitly inferred through user profiles. The AM operates on the basis of specific requests/commands coming from the Navigation and Editor Module. Figure 4 shows three adaptive views for three different users; the KB is filtered according to the permissions that characterize each users. Moreover the information available in the UP about user interests is also utilized by the AM to perform recommendation of contents and views to the users. A detailed description of the functionalities provided by the Adaptation Module is available at [21].
7) The *Navigation and Editor Module* (NEM) is devoted to visualization and editing information units and knowledge; it allows user to interactively browse the paths extracted from KB. Visualization of knowledge and information is a fundamental aspect of our model. One central reason for this is that visualization exploits several features of the human cognitive processing system. Knowledge visualization may help users to organize and reorganize, structure and restructure, assess, evaluate, elaborate, communicate, and co-construct knowledge,

and to utilize ideas and thoughts, as well as knowledge, about relevant contents and resources [22].
There are many ways to visualize a zz-structure [17]. The most simple views of them are the so-called H-view and I-view, formally described in [20]. Two examples of views built upon the zz-structure represented in Figure 3 are shown in Figure 5 and Figure 6.

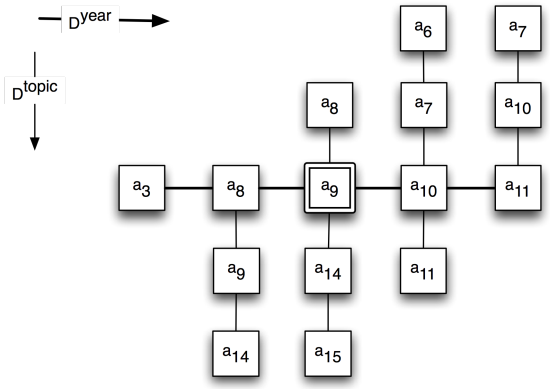

Fig. 5. A H-view with focus in a_9 (related to zz-structure of Figure 3).

Figure 5 reports the H-view, of size 5, focused on article a_9 and related to dimensions D^{year} and D^{topic}. Note that, the name H-view comes from the fact that the columns remind the vertical bars in a capital letter H. Analogously, I-view in the rows remind the horizontal bars in a capital letter I, as shown in Figure 6.

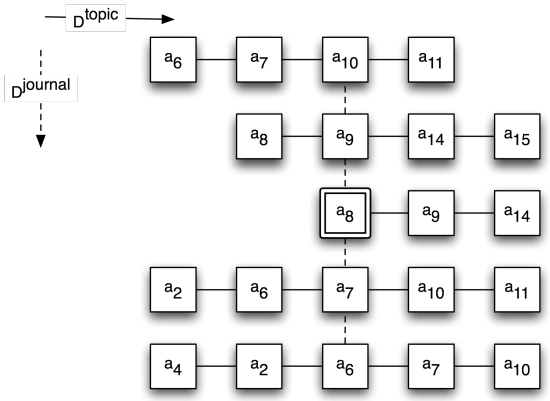

Fig. 6. An I-view with focus in a_8 (related to zz-structure of Figure 3).

The shown I-view has size 5, is focused on article a_8 and is related to dimensions D^{topic} and $D^{journal}$.

V. Conclusions and Future Work

In this paper we proposed an innovative architecture dedicated to automatically extract knowledge and meta-kowledge from traditional digital libraries in order to organize and customize the user conceptual space. Our research is on going; we plan to start experimental activities for evaluating PIE architecture in the area of medical and scientific digital libraries.

References

[1] W. P. Lee and M. H. Su, "Personalizing information services on wired and wireless networks," in *EEE*. IEEE Computer Society, 2004, pp. 263–266.

[2] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, "User profiles for personalized information access," in *The Adaptive Web*, ser. Lecture Notes in Computer Science, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds., vol. 4321. Springer, 2007, pp. 54–89.

[3] G. S. B. Nelson, "Avoiding eoverload:personalizing web content through security, eintelligence and data mining," in *Proc. of the SouthEast SAS Users Group Conference*, New Orleans, Louisiana, August 19-22 2001.

[4] P. Brusilovsky and C. Tasso, "Preface to special issue on user modeling for web information retrieval," *User Model. User-Adapt. Interact.*, vol. 14, no. 2-3, pp. 147–157, 2004.

[5] A. McCallum, "Information extraction: distilling structured data from unstructured text," *ACM Queue*, vol. 3, no. 9, pp. 48–57, 2005.

[6] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the informational, navigational, and transactional intent of web queries," *Inf. Process. Manage.*, vol. 44, no. 3, pp. 1251–1266, 2008.

[7] B. H.Karanikas, C.Tjortjis, "An approach to text mining using information extraction," in *the Proc. The Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), Workshop on Knowledge Management Theory and Applications(KMTA)*, Lyon, France, September 13-16 2000.

[8] M. Kayed and K. F. Shaalan, "A survey of web information extraction systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1411–1428, October 2006. [Online]. Available: http://portal.acm.org/citation.cfm?id=1159162.1159300

[9] K. Kaiser and S. Miksch, "Information extraction. a survey," Technical Report,Asgaard-TR, May 2005.

[10] J. Cowie and Y. Wilks, "Information extraction," in *Handbook of Natural Language Processing*, R. Dale, H. Moisl, and H. Somers, Eds. Marcel Dekker, New York, 2000.

[11] A. Micarelli, F. Gasparetti, F. Sciarrone, and S. Gauch, "Personalized search on the world wide web," in *The Adaptive Web*, ser. Lecture Notes in Computer Science, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds., vol. 4321. Springer, 2007, pp. 195–230.

[12] A. Dengel, C. Wenzel, and M. Junker, "Profile-based information supply from text sources," in *4th Multiconference on systemics, Cybernetics and Informatics and 6th International Conference on Information Systems, Analysis and Synthesis SCI/ISAS*, Orlando, FL, USA, July 2000, pp. 288–293.

[13] F. Chen, A. Doan, J. Yang, and R. Ramakrishnan, "Efficient information extraction over evolving text data," in *24th International Conference on Data Engineering,(ICDE)*, Cancun, Mexico, April 2008.

[14] E. Agichtein, "Web information extraction and user modeling: Towards closing the gap," *IEEE Data Eng. Bull.*, vol. 29, no. 4, pp. 37–44, 2006.

[15] P. Omero, N. Polesello, and C. Tasso, "Personalized intelligent information services within an online digital library for medicine: the bibliomed system," in *IRCDL*, M. Agosti, F. Esposito, and C. Thanos, Eds. DELOS: a Network of Excellence on Digital Libraries, 2007, pp. 46–51.

[16] M. Minio and C. Tasso, "User modeling for information filtering on internet services: Exploiting an extended version of the umt shell," in *UM for Information Filtering on the WWW, 5th UM Inter. Conf.*, Hawaii, June 2-5, 1996.

[17] T. H. Nelson, "A cosmology for a different computer universe: Data model, mechanisms, virtual machine and visualization infrastructure," *J. Digit. Inf.*, vol. 5, no. 1, 2004.

[18] S. Canazza and A. Dattolo, "Open, dynamic electronic editions of multidimensional documents," in *Proceedings of the IASTED European Conference on Internet and Multimedia Systems and Applications - EuroIMSA 2007*, Chamonix, France, March 14-16, 2007, pp. 230–235.

[19] A. Dattolo and F. Luccio, "A new actor-based structure for distributed systems," in *IEEE Conference on Hypermedia and Grid Systems HGS*, Opatija, Adriatic Coast, Croatia, May 21 - 25 2007, pp. 195 – 201.

[20] ——, "Formalizing a model to represent and visualize concept spaces in e-learning enviroments," in *Proceedings of the 4th Webist International Conference*, Funchal, Madeira, Portugal, May 4-7 2008, pp. 339 – 346.

[21] P. Casoto, A. Dattolo, F. Ferrara, P. Omero, N. Pudota, and C. Tasso, "Generating and sharing personal information spaces," in *Workshop on Adaptation for the Social Web - Adaptive Hypermedia 2008, Hannover, Germany, 29 July 2008. In publishing*.

[22] S.-O. Tergan and T. Keller, Eds., *Knowledge and Information Visualization, Searching for Synergies [outcome of a workshop held in Tübingen, Germany, May 2004].*, ser. Lecture Notes in Computer Science, vol. 3426. Springer, 2005.

[23] P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds., *The Adaptive Web, Methods and Strategies of Web Personalization*, ser. Lecture Notes in Computer Science, vol. 4321. Springer, 2007.