Towards the Novel Classification Schemes in Digital Libraries

BRANKO HORVAT, MILAN OJSTERŠEK
Faculty of Electrical Engineering and Computer Science
University of Maribor
Smetanova ulica 17, SI-2000 Maribor
SLOVENIA

Abstract: - Nowadays, in almost all libraries, librarians still maintain a rather obsolete practice of setting their books and other material according to a variation of the well-known UDC arrangement. However, the habits of today's users force libraries to provide a different approach for accessing library material, it should be in a digital form and easy to find. On the one hand, the library items should have richer description than mere UDC, i.e. using the automatic text indexing as well as considering the intentional point of view. On the other hand, users' needs should be specified to facilitate obtaining and delivery of the relevant items to the right users. Moreover, otherwise valuable implicit knowledge could be partially captured in a step-by-step form via forum discussions.

Key-Words: - Digital Library, Classification Schemes, Ontologies, User Profiles, Knowledge, UDC

1 Introduction

It is obvious that libraries have lost the role of the exclusive knowledge provider for doing seminar works, writing papers, preparing thesis, etc. Namely, the Internet sources and services have taken away a great share from them. People use libraries mainly as a reading-room, to use some rare dictionaries or reference books, to read daily papers, etc. The most up-to-date sources of every kind are on the Internet. There are several encyclopaedias, definition sources, how-to portals, which are very useful and users can obtain it from the home computer. Moreover, many of the sources are free of charge.

Libraries, in particular the special ones, are getting aware of the shift in users' behaviour and are trying to adapt and provide new services. First of all, they are trying to increase their amount of digital items (mostly full-text documents). They achieve it mainly via electronic subscriptions to periodicals, collecting internally issued documents, and, in lesser extent, via scanning of important paper documents, and obtaining free or payable e-books. We can also notice increased efforts of the national institutes world-wide to bring their cultural heritage into the digital form. For major nations, it is typical to bring their items selectively into the digital form due to enormous amount of items and costs, whereas, for smaller nations, an unselective transformation is common.

The majority of users favour keyword-like searching to find relevant material and few of them still browse among shelves or even search via electronic catalogue. The modern libraries are thus making efforts to provide good querying interfaces and access to the relevant material from their home computers without visiting the library. Thus, a novel arrangement (or rather novel categorizations) of the material is needed. The modern library should play a role of a knowledge supplier and a knowledge assistant for targeted users of the organization.

Users do not search only for the items with relevant topic. Moreover, they also need the e-material with an appropriate level of difficulty. Sometimes, a user wants the introductory level of the subject matter; at other occasions, some users need a detailed source of the problem. In addition, the arrangement of the e-material must be provided beforehand to find other relevant material (for instance for e-learning environment) due to performance issues.

Let us mention two well-known applications offering items and providing recommendation. CiteSeer.IST [1] is a special DL Web portal providing searching capabilities for scientific literature considering indexing and citation. It offers two type of searching, on documents and on citations. Amazon.com [2] is an e-commerce Web portal offering variety of catalogue items. It is using collaborative filtering to provide personalized recommendation.

In this paper, first, a spectrum of digital libraries is provided in section 2. Next, in section 3, the notion of a tighter collaboration between librarians and users are presented. Then, automatic and semi-automatic approaches are described in section 4 and 5, respectively. At the end, a discussion concludes the article.

2 From Items to Content and Semantics

In this section, we will provide our notions about the digital libraries spectrum from the basic level toward the advanced digital library (hence DL) functionality as depicted in fig. 1(b) (consider the vertical arrow presenting the simplest DL on the bottom and the most powerful one on the top). First, the simplest DL does not provide inherent support for content processing. In this case, the DL is in the role of the ordinary library catalogue systems with the additional feature, i.e. providing access to the full-text. Second, the next level DL does not include any additional mechanisms for the knowledge management since only the full-text search is available. However, this capability of the DL is a major step forward comparing to the mere catalogue systems. For the lower scalability requirements, a usual data base or a file systems capability features can be utilized. For higher performance, a proprietary solution should be provided.

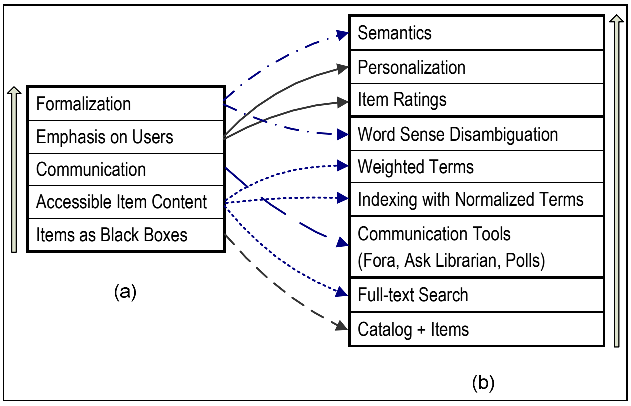

Fig. 1. A coarse (a) and a fine stack (b) of knowledge support spectrum for a DL.

Third, the DL provides also communication tools, e.g., domain forums, ask librarian, polls, etc. They can serve as a rich source of the procedural knowledge. Furthermore, they enable the institution to acquire the explicit knowledge out of the employees' implicit one. Of course, they have to be maintained by a moderator. So, a librarian is in charge of watching development of the discussions, cleaning the content, and attaching an appropriate e-material to them. This way, the forum discussions are enriched with e-items, and vice versa. Fourth, the full-text search can be accompanied by the text indexing with the term normalization. Only the most informative terms (normally words) are selected for a document representation (indexing). The terms are transformed into their base form called lemma. This task is called lemmatization. Fifth, the informative terms are joined with the score of the term's information share within the document according to the document's remaining keywords. On the basis of the efficient data structures for the terms-to-documents and for the documents-to-terms representation, the search of the most relevant (or at least similar) documents is provided; whether upon the search query or another document.

Sixth, eliminating some natural language ambiguity can lead the library to effective solution of the polysemy and the synonymity problem. With the help of thesauri and other linguistic tools, a more neutral content description can be provided. For instance, the synonyms can be added beside the occurred terms. Moreover, the content can be represented even by the meaning instead of the synonyms. In the first case, also the queries have to be extended with the synonyms. In the latter case, beside the terms, also the context must be taken into account. This process is called word sense disambiguation (WSD).

Seventh, we can further expand our DL with the inclusion of the user members and their wishes. Thus, at the selection of the e-material, not only automatically obtained data about document properties are considered, but also opinions of the users about certain e-items are taken into consideration. Beside the explicit ratings, also estimated implicit ratings can be employed. In the first case, a user is able to rate in the specified range, for instance from 1 to 5. In the latter case, we acquire rates in a more sophisticated way, e.g., via usage analysis (page views and time spent) or by short targeted questioning about the e-item, which must give an impression of a by-the-way style. Eighth, the library enables the individual profiles (models) of their users. This way, we can provide the personalization of the search for e-material. Namely, in the result set, we place higher those matched items, that have high rate considering the content features of the user, or according to the rates of other users, which are the most similar to that user. Besides, we can provide automatic notification and even delivery of the appropriate items to the users. Ninth, if we provide the DL with more semantic level, we get the library, which is based on the semantic Web portal. In this case, we have ontologically described documents, the structure of the DL, domain knowledge, etc. The first helps us to better describe a document other than by plain key terms set. The second presents the possibility to provide the current location inside the DL and enables easier navigation. The third incorporates each single document into the broader context.

Above, we have tried to present levels of the DL features beginning with the simplest DL and, gradually, coming to the most demanding one focusing on knowledge. Of course, the order can be a matter of discussion. It is certainly true that the personalization is not a precondition for using formalized knowledge or WSD. However, it still represents some natural order considering technology and the structure of employees.

Let us view some more coarse conception of it now as shown in Fig. 1-a. There is also the vertical arrow here, presenting progression of DL categories. Moreover, relationships between coarse and fine concepts are provided (Fig. 1 a and b, respectively). First, no full-text processing is provided. There is only the link to the e-item with the corresponding metadata. Second, full-text functionality is included, as for instance full-text search, indexing, and other non-semantic approaches. Third, communication facilities like forums are used. Fourth, there is emphasis on the individuals and his features. Fifth, approaches for WSD and advanced formalization of the meaning are applied.

In our opinion, at least indexing as inner text processing should be provided for every DL. We recommend basic communication tools as polls and forums. Also recommendation and personalized search systems as well as user notification are advised. An additional tool as a thesaurus (possibly in the ontological form) or a statistical tool is necessary to perform natural language processing (NLP), as WSD. Additional skills are necessary for the librarians in order to manage knowledge formalization, i.e. descriptions via ontologies, which must include understanding ontologies (knowledge representation via entities and relationships between them) as well as applying tools for building them, and, after all, also understanding existing domain ontologies (knowledge); whether for certain field within library profession or for the fields covered by the library.

3 Synergy between Users and Librarians

In our system, DL users as well as librarians should have much more active role comparing to classical libraries. Normally, a user visits a library and searches the catalogue to find appropriate books. After brief inspection, she makes a final selection from the available ones and a librarian registers this in the system. One of the main tasks of librarians is also obtaining new items and registering them into the catalogue. Unfortunately, they have no insight into what items are really needed.

We think both parties should work with more synergy. Libraries, in particular the special ones, have common goals with their users as depicted in Fig. 2. Users seek items to accomplish their tasks or because they are interested and librarians would like to provide good knowledge services to make the user satisfied and the institution competitive. They both only have to utilize the available technology. The user must actively contribute her share. She must specify her goals, wishes, etc. Among others, for a user, her current educational activity should be provided (e.g. working on his MSc degree), which field she studies, what languages she speaks, etc. It is invaluable to know, which items she has already read, how she characterizes them, and to specify those she would need but are unavailable. Users should be encouraged to provide scores to the read items with accompanying comment. A librarian actually could play a role of the second mentor beside the main expert mentor. On the other hand, the librarians analyse user needs and actively obtain items to fulfil users' needs. As soon as the library gets them the interested users are notified with suitable items.

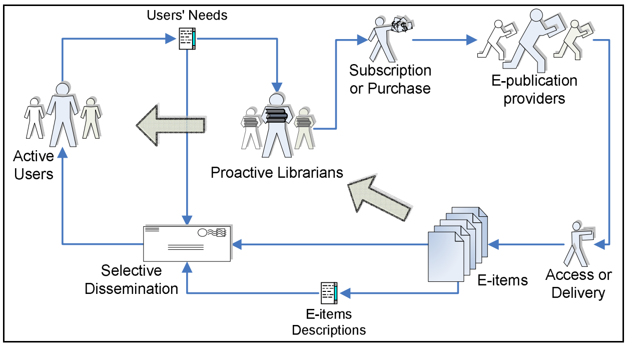

Fig. 2. Librarians, who are informed about users' needs, obtain the relevant items and deliver them to the relevant users.

3.1 Proactive Librarians

The classical librarian has much less responsibility than the employee of DL will have. Modern librarians will work more in the sense of information system co-creator. They will complement existing categorization schemas and develop new ones (Fig. 3). The institution domains will be evenly assigned among them; thus, librarians will become specialists for particular domains. Librarians will be the best informed team about the current activities of organizational entities as well as of employed individuals and other parties (students). Their special work will be to encourage internal issuing documents and discussions to acquire implicit knowledge and correct registration of them in the system. This way, all thesis and diplomas will be much more accessible and available for interesting readers. They will also be effectively categorized and the relationships between items could be inferred. The evidence of knowledge and activities as well as needs of individual users will come to the front.

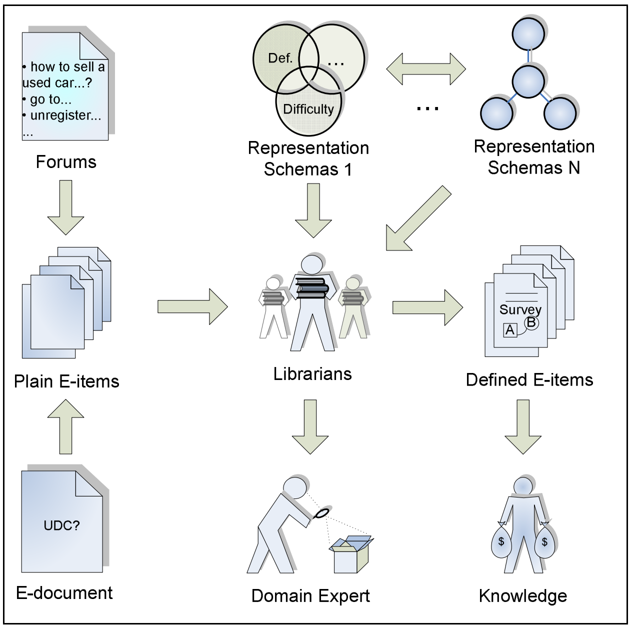

Fig. 3. Librarians obtain needed items and define them.

3.2 Active Users and Their Profile

Some Web portals, as DL, having rich collection of items and a lot of users should provide some personalization features. A user profile is necessary to enable effective personalization (automatic tailoring of content and appearance) as shown in [4] and [5]. On the one hand, each user must be considered as an individual and, on the other hand, as a part of one of the user groups (segments) in order to be provided with targeted services. The first is necessary to accomplish content-based personalization and the second for the social-based personalization. A user profile usually stands for user features obtained over several visits. If no user identification is possible then personalization on current visits is employed still using previous unidentified visits information. Beside the data based on the usage, for a profile, it is normally very important to have basic demographics data, such as gender, age, etc. Moreover, a good source of information are ratings, whether implicit or explicit ones, and poll data. On that basis, we can find similar users, segment them, and create recommendations not seen yet. Besides, each user can have its own set of most important keywords, which can be further reduced into concepts as presented in [6] and [7]. Also, user-defined preferences should be considered that can be obtained via a questionnaire filled out on registration. They should include demographics data, language skills, and position (rank) within the institution.

In addition, forum discussions should be considered that a user has taken part in. It is important to infer whether a user was putting a question and thus searching a solution to a problem, or he/she proposes a solution [3]. This way, we can partially capture implicit knowledge of employees and thus of institution. It must be clear what the user's current educational and other efforts are in order to obtain help from DL; whether a user addresses some problem from the engineer, management (organizational, economical), or theoretic point of view etc. Actually, every user must have provided the map of knowledge, read items, and intentions (high school, diploma, and other certificates fields). The most important thing here is that a user specifies his future wishes and waits until something relevant comes.

3.3 Definition of Items

As we have already emphasized in the title, we would like to transcend the UDC realm and come to more informative and useful definition of a document (or generally e-item). The UDC descriptor tells nothing more than mere subject field and possibly additional (alternative) ones. Even librarians realize it has more or less historical and statistical value and maybe partially for arranging books on the shelves to help librarians or visitors with finding books. An additional categorization and description should be provided. Moreover, the best solution is to use several of them in parallel as a kind of a multidimensional categorization. Beside the UDC, there are several other taxonomies as ACM in [8], etc. The idea is not so much how to provide several taxonomies of the professional fields but more how to utilize categorization for different purposes.

First, the institutional taxonomy of topics should be considered. In the case of the university library, an e-item can be linked with the courses hierarchy to represent a relationship to lectured subjects. This is of great importance for lecturers and in particular for students. However, this has little meaning for researchers, which have no role in educational process. We can say this feature or this scheme covers student and lecturer intentional viewpoint in the sense how each e-item contributes to a particular course subject.

Second, we propose characterization of the e-items in the sense of usefulness. For instance, an e-item can be an exhaustive source of corresponding definitions; or can be a good introduction, maybe even in the form of presentation slides; it can be a survey serving as a good contextual reading; or it can have an advanced, detailed, expert, and profound character. Of course we can use combination of them if necessary. Third, the difficulty level should be provided, e.g. popular, primary or high school level, undergraduate, postgraduate or specialist study, etc.

Fourth, relationships between e-items, i.e. inter-relationships should be determined. For the specific item, a required reading can be specified, i.e. the items that could play an introductory role. On the other hand, recommended further reading specified or even inferred from the "introductory" items (these two relationships can be, in fact, inverse). Beside the concrete items, it can be presented also in the sense of knowledge required (mathematics on matrix, probability calculus, etc.). Fifth, a special care must be paid for internally created documents. It must be obvious if it is a draft, final, or even peer-reviewed work. This can be called the status of the document.

Several next topics cover indexing-based features. Sixth, we propose a measure to assess the diversity of the document (the number of concepts covered in the e-item). It is based on the structure as well as on the content. Seventh, also the basic term ontology can be built whether general but more possibly domain specific. Eighth, link to particular forum discussion or FAQ can be provided to present a broader solution for a specific problem. Last but not least, the language and the type of the e-item should be also provided.

Some things described here could be done automatically, some semi-automatically, and the other manually. Let us have a look at all these kinds of approaches.

4 Automatic Approaches

The aim of information systems, thus also of DL, is to facilitate or completely remove tedious work from the librarians and users. At the same time, it is an opportunity to provide advanced services otherwise impossible by the current human resources. There are several sources and ways to obtain data and to accomplish tasks automatically. First, each text-based document and other multimedia item could be at least indexed with weighted terms. Second, the next level could be extracting concepts (by LSI – latent semantic indexing or CI – concept indexing as in [6] and [7]). Third, the explaining introductory references could be extracted (from the introduction section). Fourth, similar and possibly relevant documents could be provided based on the specified items or query string.

Fifth, the base literature could be identified for specific field via curriculum course specification. Each subject within a course has a reference section to specify basic sources for students that cover the subject matter. The sources there are normally provided in the descending order according to the usefulness. All these sources, especially the first one, can be acquired as a basic reading for the subject; and, in some extent, also for the e-items that are categorized into such class. It is useful for instance for e-learning purposes. Sixth, in some cases, a user and his current educational activity can be automatically spotted via institutional information systems with the corresponding study programme and selected courses.

5 Manual Decisions

There are tasks that could not be accomplished automatically without user intervention. However, even for such tasks, computer assistance could be provided. There is no current idea how to provide help with defining difficulty or usefulness levels. It is very important to employ authors of the items as much as possible. For instance, a student passing a diploma should help categorize and describe the work. Normally, the title is the best basis for the beginning of building term ontology of the item. The next step could be summary. It is important to identify also future propositions of the author. For sure, the task of presenting institutional taxonomy must be done by hand, however it is a one time task. Each laboratory should also provide own research ontology.

6 Discussion

A special attention must be paid for final works as diplomas, master theses, etc. Namely, they provide qualitative overview, and other references. Also, the mentors for these works must be considered as well as conferences and other community meetings arranged within the institution. The same can be done for seminal works etc. We could use several presentations and other papers as a source of analysis, e.g. for extracting references in introductory section, employing references in general, etc.

7 Conclusion

In this paper, a notion is presented of shifting the activity from a user to a librarian to provide the user with targeted services and instant notification of new relevant material in DL. An idea of connecting material as well as communication records is given to provide organized institutional knowledge and human resources. A proposition of several viewpoint categorization is provided to better describe each e-item. Advice about automatic acquisition sources of some data in the case of educational institution are mentioned.

References

[1] S. Lawrence, C. L. Giles, K. Bollacker, Digital libraries and autonomous citation indexing, *IEEE Computer*, Vol. 32, No. 6, pp. 67-71, 1999.

[2] G. Linden, B. Smith, J. York, Amazon.com recommendations: item-to-item collaborative filtering, *IEEE Internet Computing*, Vol. 7, No. 1, pp. 76-80, 2003.

[3] N. Matsumura, D. E. Goldberg, X. Llora, Communication gap management for fertile community, in press, *Soft computing*, *Springer*, 2006.

[4] G. Adomavicius, A. Tuzhilin, Using data mining methods to build customer profiles, *IEEE Computer,* Vol. 34, No. 2, 2001

[5] B. Mobasher, H. Dai, T. Luo, Y. Sun, J. Zhu, Integrating Web Usage and Content Mining for More Effective Personalization, *Proceedings of the International Conference on E-Commerce and Web Technologies* (ECWeb2000), Greenwich, UK, 2000.

[6] I. S. Dhillon, D. S. Modha, Concept decompositions for large sparse text data using clustering, *Kluwer Academic Publishers,* 2000.

[7] C. H. Papadimitriou, P. Raghavan, H. Tamaki, Latent Semantic Indexing: A Probabilistic Analysis, Symposium on Principles of Database Systems, *ACM Press*, 1997.

[8] Top-Level Categories for the ACM Taxonomy: www.computer.org/portal/pages/ieeecs/publications/author/ACMtaxonomy.html, visited July 2006.