Towards knowledge-based digital libraries
Feng, L.; Jeusfeld, M.A.; Hoppenbrouwers, J.J.A.C.

Published in:
SIGMOD Record: A quarterly publication of the ACM Special Interest Group on Management of Data

Publication date:
2001

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 25. Jan. 2020
Towards Knowledge-Based Digital Libraries

Ling Feng Manfred A. Jeusfeld Jeroen Hoppenbrouwers
InfoLab, Dept. of Information Management, Tilburg University
PO Box 90153, 5000 LE Tilburg, Netherlands
{ling, jeusfeld, hoppie}@kub.nl

Abstract

From the standpoint of satisfying human’s information needs, the current digital library (DL) systems suffer from the following two shortcomings: (i) inadequate high-level cognition support; (ii) inadequate knowledge sharing facilities. In this article, we introduce a two-layered digital library architecture to support different levels of human cognitive acts. The model moves beyond simple information searching and browsing across multiple repositories, to inquiry of knowledge about the contents of digital libraries. To address users’ high-order cognitive requests, we propose an information space consisting of a knowledge subspace and a document subspace. We extend the traditional indexing and searching schema of digital libraries from keyword-based to knowledge-based by adding knowledge to the documents into the DL information space. The distinguished features of such enhanced DL systems in comparison with the traditional knowledge-based systems are also discussed.

1 Introduction

With the exponential growth of information in the Web, there is a demand for new evolutionary technology to support effective search and indexing functionalities. Digital libraries are good examples to investigate new approaches to effective use of large information repositories because of the long tradition of conventional libraries in supporting human’s information needs. They integrate a variety of information technologies which provide opportunities to assemble, organize and access large volumes of information from multiple repositories, while making distributed heterogeneous resources spread across the network appear to be a single uniform federated source [SC99]. Under the assistance of DL systems, users can move from source to source, seeking and linking information automatically or semi-automatically. From a user’s perspective, DL systems establish a fundamental infrastructure for a bulk of digital information and services associated with users’ information acts.

Traditionally, when people retrieve information, their activities are classified as searching or browsing [CDMS94, Hop98]. Searching implies that the user knows exactly what to look for, while browsing should assist users navigating among correlated searchable terms to look for something new or interesting. So far, most of the major work on DL systems falls into these two categories. DL research has neglected to support systematic acquisition of knowledge about the DL content. This has been the role of a traditional librarian who could direct users to the right articles when asked for advice. Our goal is to establish this role by an electronic counterpart. The content of its knowledge base is created in a collaborative effort.

1.1 Related Work

To support efficient searching activity, efforts have been made in developing retrieval models, building document and index spaces, extending and refining queries for DLs [FBY92, CLvRC98]. In [DvR93], index terms are automatically extracted from documents and a vector-space paradigm is exploited to measure the matching degrees between queries and documents. Indexes and metadata can also be manually created from which semantic relationships are captured [BS95, Dao98]. Furthermore, the information space consisting of a large collection of documents can be semantically partitioned into different clusters, so that queries can be evaluated against relevant clusters [Wil88]. According to topic areas, a distributed semantic framework is proposed in [PH99, Mil00] to contextualize the entire collection of documents for efficient large-scale searching. To improve query recall and precision, several query expansion and refinement techniques based on relational lexicons/thesauri or relevance feedback have been explored [VWSG97, Eft93, JGR+95]. A recent work incorporates knowledge about the document structures into information retrieval, and the presented query language allows the assignment of structural roles to individual query terms [WFC00].

Since one DL usually contains lots of distributed and heterogeneous repositories which may be autonomously managed by different organizations, in order to facilitate users’...


browsing activities across diverse sources easily, many efforts have been engaged in handling various structural and semantics variations and providing users with a coherent view of a massive amount of information. Nowadays, the interoperability problem has sparked vigorous discussions in the DL community [SC99, SMC+99, Sch98, Sch95, Che99, PBL99, PCGMW98]. The concept extraction, mapping and switching techniques, developed in [BHCS99, MG95, CSN97], enable users in a certain area to easily search the specialized terminology of another area. A dynamic mediator infrastructure [MGMP00] allows mediators to be composed from a set of modules, each implementing a particular mediation function, such as protocol translation, query translation, or result merging [PBJ+00]. [PL99, JL97] present an extensible digital object and repository architecture FEDORA, which can support the aggregation of mixed distributed data into complex objects, and associate multiple content disseminations with these objects. [KW95, PCD+98, PH96] employ the distributed object technology to cope with interoperability among heterogeneous resources. With XML becoming the Web data exchange standard, considerable work on modeling, querying and managing semistructured data and non-standard data formats are conducted to enable the integration of heterogeneous resources [DBJ99, MW99, BDT99]. The experiences in constructing DL archival repositories, user interfaces, and cross-access mechanisms, etc. are extensively described in [HP00, CGM99, CCM00, HBS96, Hou95, PW97, Liu99, CMFH99, Kan98, BSH97, BL97]

1.2 The Inadequacy

Despite lots of fruitful work in the digital library area, from the standpoint of satisfying human’s information needs, the current DL systems suffer from the following two shortcomings.

Inadequate High-Level Cognition Support

The traditional use of DLs is keyword-based. Users request information by entering some keywords, and DL systems return matching documents. But users expect more than this. Typically, users have some pre-conceived hypotheses or domain-specific knowledge. They may desire the library to confirm/deny their existing hypotheses, or to check whether there are some exceptional/contradictory documental evidences against the pre-existing notions, or to provide some predictive information so that they can take effective actions. For example, a user working in a flood-precaution office is concerned about whether there will be floods in the coming summer. According to his/her previous experience, it seems that “a wet winter may cause floods in summer”. In this situation, instead of using disperse keywords to ask for documents, the user would prefer to pose a direct question to DLs like “Does a wet winter cause floods in summer?” and expect a confirmed/denied intelligent answer as well as a series of supporting literatures to justify the answer, rather than a list of articles lacking explanatory semantics and waiting for his/her further checking.

Inadequate Knowledge Sharing Facilities

Traditional libraries are a public place where a large extent of mutual learning, knowledge sharing and exchange can happen. A user may ask a librarian for searching assistance. Librarians may collaborate in the process of managing, organizing and disseminating information. Users themselves may communicate and help each other in using library resources. When we progress from physical libraries to virtual DLs, these valuable features must be retained. Future DLs should not just be simple storage and archival systems. To be successful, DLs must become a knowledge place for a wide spread of knowledge inquiry, sharing and propagation. In the above example, if the DL makes readily available knowledge and expertise to the public, users can save the effort on time-consuming searching and consultation with librarians and/or experts. The working effectiveness and efficiency can thus be improved. Also, as machine knowledge does not deteriorate with time as that human knowledge does, for long-term retention, DL systems offer ideal repositories of the knowledge in the world. Unfortunately, such a knowledge sharing function of DLs have not received extensive exploration so far.

2 A Two-Layered DL Cognitive Function Model

We categorize users’ behavior on the use of DLs into low-level cognitive act and high-level cognitive act. Figure 1 illustrates a proposed two-layered DL cognitive function model to support different levels of users’ cognitive requests.

![Figure 1: A two-layered DL cognitive function model](image)

2.1 Low-Level Cognition Support

We view traditional information searching and browsing as low-level cognitive acts.
Searching. The target of searching is towards certain specific documents. One searching example is “Look for the article written by John Brown in the proceedings of VLDB ’88.” As the user’s request can be precisely stated beforehand, identifying the target repository where the requested document is located is relatively easy. Primarily, the ability to search indexes of repositories can support the searching activities.

Browsing. Different from searching whose objective is well-defined, browsing aims to provide users with a conceptual map, so that users can navigate among correlated items to hopefully find some potentially useful documents, or to formulate a more precise retrieval request further. For instance, a user reads an article talking about a water reservoir construction plan in a certain region. He/she may want to know the possible influence on ecological balance. By following semantic links for the water reservoir plan in the DL, he/she navigates to the related “ecological protection” theme, under which a set of searchable terms with relevant documents are listed for selection.

To facilitate browsing, DLs must integrate diverse repositories to provide users with a uniform searching and retrieval interface to a coherent collection of materials. The capability that enables navigation among a network of inter-related concepts, plus the searching capability on each individual repository, constitute the fundamental support to browsing activities.

2.2 High-Level Cognition Support

In contrast to the low-level cognition support whose eventual goals are documents, the high-level cognition support layer can provide not only documents but also knowledge-level answers to human’s high-order cognitive questions, together with documental justifications and evidences. For example, in response to the high-order cognitive requests like Q1: “Does wet winter cause floods in summer?” Q2: “Give me articles which talk about the cause of summer floods.” Q3: “Give me articles which talk about the influences of wet winter.” it is desirable for DL systems to provide question-answering, as well as relevant justifications for holding the answers. For example, the justifications for Q1, Q2 and Q3 will consist of a series of articles talking about “wet winter causes floods in summer.”

The provision of high-level cognition support adds values to DLs beyond simply providing document access. It reinforces the exploration and utilization of information in DLs, and advocates a more close and powerful interaction between users and DL systems. With this high-order cognition assistance, ordinary people will be able to find things to solve their real information problems themselves. From the aspect of DL systems, to realize such a high-level cognition function, substantial information analysis needs to be done. This inevitably involves the navigation and cross-correlation of information items across multiple repositories in DLs, and acquisition of intelligent knowledge in answering users’ high-level cognitive questions.

3 An Enlarged DL Information Space

To provide high-order cognition support, we further develop a DL information space consisting of two component subspaces, namely, knowledge subspace and document subspace, as shown in Figure 2.

![Figure 2: A DL information space for high-order cognition support](image)

3.1 The Knowledge Subspace

The basic constituents of the knowledge subspace are knowledge, such as hypotheses, rules, beliefs, etc. In this initial study, we focus on hypothesis knowledge coming from domain experts in empirical science. Each piece of hypothesis describes a certain relationship among a set of concepts. For example, the hypothesis “H: Wet winter may cause summer floods” explicates a causal relationship between a cause “C1: wet winter” and the effect “C2: summer flood” it has.

Here, we use a predicate which takes a set of concept terms as arguments to represent each hypothesis. A concept term can be either an atomic concept or a composite concept. Atomic concepts are the building blocks of sentences (e.g., “dog”, “animal”, “traffic-jam”, “wet-winter”, “summer-flood”, etc.), conveying the most fundamental cognitive knowledge in human society, while composite concepts are built up from atomic concepts through the concept conjunctive operator (\(\cap\)). For example, “warm-winter \(\cap\) wet-winter” is a composite concept. At the moment, we focus our study on binary predicates associated with two concept terms: a left-side concept term and a right-side one. For example, the hypothesis “Wet winter may cause summer floods” can be expressed as Cause (“wet-winter”, “summer-flood”). “Air pollution may cause acid rain and hot-weather” is another hypothesis example which can be described as Cause (“air-pollution”, “acid-rain \(\cap\) hot-weather”).

Based on different concept relations (e.g., is-a, part-whole, synonym, and antonym, etc.) defined in the concept base, we can correlate relevant hypotheses and formulate a hypothesis lattice around one theme. For example, a more general
hypothesis in respect to \( H \) is like “\( H' \): wet winter may cause river behavior”, as “summer-flood” is a more specific concept term compared to “river-behavior”.

The knowledge subspace of a DL is thus made up of a number of hypothesis lattices in different domains.

### 3.2 The Document Subspace

Under each hypothesis is a justification set, giving reasons and evidences for the knowledge. These justifications, comprised of articles, reports, data, etc., constitute the document subspace of the DL information space. In Figure 2, we have a set of supporting articles for hypothesis \( H \), which comment that “wet winter is an indicator of summer floods”.

It is worth notice here that the document subspace challenges traditional DLs on literature organization, classification, and management. For belief justifications, we must extend the classical keyword-based index schema, which is mainly used for information searching and browsing purposes, to knowledge-based index schema, so that information in DLs can be easily retrieved by both keywords and knowledge.

### 3.3 Linking the Two Subspaces

The knowledge subspace (i.e., the collection of hypotheses) subsumes a wide range of knowledge coming from human experts in different areas. Each piece of knowledge in the knowledge subspace is linked to a set of justification documents in the document subspace. The linkage between the two subspaces can be built in a number of ways: 1) Experts indicate relevant documents while inputting the knowledge; 2) DL systems perform keyword-based searching. From the results obtained, relevant justification documents are filtered by either experts or DL systems through a more close examination of the documents. 3) DL users, who find justifications for certain knowledge, mark the corresponding documents. Later, other users can re-use these findings.

### 4 Discussions

Although applications of artificial intelligence to library science have been extensively investigated in the literature, and many library-oriented expert systems have been developed, most of them essentially aid in carrying out the support operations of libraries, such as descriptive cataloging, collection development, disaster planning and response, reference services, database searching, and document delivery, etc. [LS90, LS97].

In this study, we extend the traditional role of DLs as information provider to information & knowledge provider by incorporating both knowledge and documents into the DL information space. Compared to the traditional knowledge-based systems, such DL systems enhanced with knowledge elements have the following distinguished features.

1) Function. Knowledge-based systems are designed to apply logical inference rules to make judgement in processing business routines or come up with a conclusion to a certain pre-defined problem [And92]. A production rule used in knowledge-based systems usually has the format “IF \( x \) THEN \( y \)”, where the IF part is a premise and the THEN part refers to the conclusions or consequences. On the contrary, the mission of a DL system equipped with a knowledge subspace is to make expertise knowledge widely available to the public. We can view such a system as an information & knowledge dictionary, since a huge body of knowledge of various kinds in the world, together with their justification documents, is preserved, classified and maintained inside its information space. From DLs, users can obtain not only the requested documents, but also intelligent answers to their high-order cognitive questions.

2) Scope. A knowledge-based system intends to solve problems in a narrow domain, e.g., company delivery charge, heart disease diagnosis, etc. The rules stored in its knowledge base are thus limited to a particular field. Comparatively, the scope of the knowledge subspace of a DL is broad, covering a wide spread of disciplines. Users with different backgrounds can turn to DLs for expert-like helps in carrying out their work.

3) Content. With the continuing developments in storage and communication technologies, a tremendous amount of structured, semi-structured, and unstructured information assets is collected and maintained within DLs. While we extend the DL information space to incorporate knowledge, such a huge body of documents constitutes knowledge justifications for users’ further reference. However, this is not the case for traditional knowledge-based systems, which contain only a limited amount of rules and facts in a particular field of expertise.

### 5 Conclusion

Motivated by the problems - (i) inadequate high-level cognition support; (ii) inadequate knowledge sharing facilities - with the present-day digital library systems, we introduce a two-layered digital library function model to support different levels of human cognitive acts. The low-level cognition support aims to provide users with requested documents, as what information searching and browsing do, while high-level cognition support can provide not only relevant documents but also intelligent answers to users’ high-order cognitive questions, as well as a set of documental justifications. The proposed information space consisting of a knowledge subspace and a document subspace can facilitate users to solve their high-order cognitive problems.

We view this work as a first step, with a number of interesting problems and challenges remaining for future work.

(1) To facilitate high-order cognitive activities, efficient stor-
Management of the knowledge & document subspaces is very important and must be carefully planned. This demands effective indexing strategies for both knowledge and justification documents. (2) Efficient knowledge inference and navigation mechanisms must be built to support users’ question-answering. (3) A flexible and easy-to-use query language is to be designed to help DL users make the best of information and knowledge assets in solving their problems. Currently, we are researching various methods of knowledge acquisition to fill the knowledge subspace and building the links between the knowledge subspace and the document subspace. Our eventual target is to develop an enhanced DL system, which can empower human with real actionable knowledge in solving their real information problems.

References


