1

Introduction to E-Librarian Services

Throughout history, libraries have always been the carriers of knowledge and the instruments for learning from the experiences and endeavors of previous generations. More recently, digital libraries have removed the physical walls of classical libraries by making accessible tremendous amounts of knowledge in multimedia form via a broad variety of electronic tools.

This evolution of libraries confronts modern information retrieval (IR) systems with new challenges. The content in such digital repositories is getting increasingly more complex to process, as it adds multimedia content, such as audio and video documents, to its previously exclusively text-based content. This in turn renders the search of information inside digital libraries increasingly difficult.

Parallel to the changes of content, the expectations and demands which users have towards IR systems are becoming more ambitious. Users expect simple interfaces for optimized and powerful search engines which are able to find the exact answers to their questions.

1.1 From Ancient to Digital Libraries

The collection of written knowledge in some sort of repository is a practice as old as civilization itself [KK01]. About 30,000 clay tablets found in ancient Mesopotamia date back more than 5,000 years. Archeologists have uncovered papyrus scrolls from 1300 – 1200 BCE in the ancient Egyptian cities of Amarna and Thebes, and thousands of clay tablets in the palace of King Sennacherib, an Assyrian ruler from 704 – 681 BCE at Nineveh, his capital city. The name for the repository eventually became *library*, which, in Greek, refers to "a collection of books". The Latin word for *library* is *bibliotheca*.

Early collections of papyri, usually considered as archives rather than libraries, may have surfaced from the Near East, but the ancient Greeks propelled the idea through their heightened interest in literacy and intellectual

S. Linckels and C. Meinel, *E-Librarian Service: User-Friendly Semantic Search in Digital Libraries*, X.media.publishing, DOI 10.1007/978-3-642-17743-9_1,

Fig. 1.1. Artistic Rendering of "The Great Library of Alexandria" by O. Von Corven.

life. Public and private libraries flourished through a well-established process: authors wrote on a variety of subjects, *scriptoria* produced the books, and book dealers sold them. Copying books was an exacting business in high demand.

The early word for book was *codex* (Latin for: block of wood), which was a Roman invention that replaced the *scroll*. A codex is a book in the format used for modern paper-based books, with separate pages normally bound together inside a cover.

Throughout the 1600s and 1700s, libraries surged in popularity. They grew as universities developed and as national state-supported collections began to appear.

The 20^{th} century saw the continued development of the library through education and organization. Libraries in educational institutions have developed a wide range of services to meet the educational objectives of their parent institutions. School libraries clearly need to support the curriculum, but they also collect books and other materials to encourage reading and inquisitive thinking, as well as to meet the needs of the teachers and administrative staff.

Libraries started to change with the appearance of microfilms in the 1930s and the development of early electronic databases in the 1950s [MS09]. The library profession was becoming increasingly technical.

The growth in electronic media available to the general public and its ease of use have been the catalysts for librarians to adapt to the new information landscape, to develop new services, and to improve library provision. The term *digital library* was first made popular by the NSF-DARPA-NASA Digital Libraries Initiative in 1993.

In the 1990s, large-scale projects were initiated that aimed to digitize and preserve books and other paper-based documents. Improvements in optical character recognition (OCR) and new standards in electronic book formats led to initiatives like Google Books[1], Universal Digital Library[2], Project Gutenberg[3], and Internet Archive[4].

In the past few years, we have been able to witness a tremendous increase in the availability of information throughout knowledge repositories in digital form. For example, at the Hasso-Plattner-Institut (HPI) in Potsdam, Germany, over 30 hours of university lecture videos about computer science are produced every week and added to its archive, which has nearly 10,000 hours of archived lecture material. Most of it is published at the online tele-TASK archive[5].

More recently, novel forms of digital libraries have been developed and intro-duced to a wide audience on the World Wide Web (WWW), e.g., wikis, weblogs, social networks, and file sharing services. Although these types of libraries have unconventional and sometimes chaotic approaches to authoring and organizing content, they manage to attract millions of users every day. Some of these libraries are self-organizing systems where both content and structure are defined by the individuals of the community. This means that ordinary users get involved in the content creation process that was previously mostly restricted to professional information providers. The natural development of library system has thus led to a democratization of the knowledge building and archiving processes.

[1] http://www.gutenberg.org/
[2] http://www.ulib.org/
[3] http://www.gutenberg.org/
[4] http://www.archive.org/
[5] http://www.tele-task.de/

1.2 From Searching to Finding

1.2.1 Searching the Web

Web search engines have changed the way people are looking for information. Today, searching for information often means using the WWW by *googling* keywords and browsing through the first page of the results list.

How did people look for information before there was Google, Yahoo!, or even the WWW? People went to libraries and asked the librarian for assistance, or they just walked along the shelves and browsed for books which matched some information they had, like the name of an author, a title, a genre, or some other information about a document, such as the color of the cover or the year of publication.

The librarian or the customer had the possibility to use a register of library cards, which contained information about each document (see figure 1.2)[6]. Generally, a *controlled vocabulary* was used for the description of the documents, e.g., title, author, year of publication, ISBN, type of document, editor, and number of pages. This supplementary information is called *metadata*, i.e., data about data. Most important was a reference or identifier where to find the document in the library.

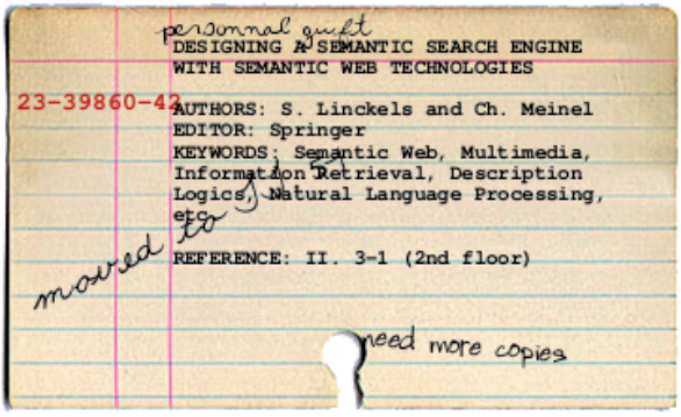

Fig. 1.2. Library card with metadata about a document.

Computer-assisted search of information started in the 1960s with the adoption of database management systems and applications. Classical IR is based on a repository of indexed documents and information needs from a user. The latter are translated into database queries. Different models were created to find best matching documents, e.g., Boolean model or vector space model (see chapter 5).

[6] Image created at: `http://www.blyberg.net/card-generator/`

The WWW can be perceived as an enormous distributed multimedia database with billions of *hypertext* documents. The WWW uses *hyperlinks* as supplementary help for the navigation. The first Web search engine was *Aliweb*, which appeared in 1993 and still runs today [Kos94].

First generation search engines (1995 – 1997) used almost exclusively on-page data such as text and formatting information to compute and rank the answer set. Second generation search engines (since 1998) like Google use off-page, Web-related data, such as link analysis, anchor-texts, and click-through data. The third generation of search engines (since 2003) try to blend data from multiple, heterogeneous sources, trying to answer "the need behind the query". The computed results are customized according to the user's needs and take into account the user's personal data background, context, and intention. They include social networking information, tagging, user feedback, semantic analysis, recommendations, and trustworthiness of information.

Similar to classical IR systems, Web search engines work by indexing large number of Web pages. *Web crawlers*, also called "spiders", "robots", or "agents", automatically retrieve Web pages and follow every link they encounter. The content of each Web page is then analyzed to determine how it should be indexed, e.g., words are extracted from the titles, headings, or meta-tags.

When a user enters a query into a search engine, typically by using keywords, it browses its index and provides a list of best-matching Web pages. Most search engines support the use of Boolean operators *and*, *or*, and *not* to further specify the search query.

Search engines have methods to rank the results and to yield the best results first. How a search engine decides which pages are best matches and in what order the results are listed varies widely from one system to another.

While the WWW grows at an increasing pace, its users and content change as well. The Web is no longer only about computer science, routers, and servers. The Web is about people, information, and entertainment; it turns into a *social Web*. This new Web, with its blogs, wikis, and social networks, opens up entirely new opportunities for novel applications and entrepreneurs, and simultaneously introduces new challenges to Web designers and search engines.

Classical keyword-based search engines rely on the fact that the user enters a part of the answer. For example, if you want to know what the tasks of a network protocol are and if you simply enter keywords like "protocol" and "task", then a search engine might yield documents about protocols and tasks, but not necessarily only documents which explain the tasks of protocols. If you knew a part of the answer, e.g., that error-handling is a task of a protocol, you could add the keyword "error-handling" to the query in order to narrow down the search.

1.2.2 Searching Multimedia Knowledge Bases

Multimedia information retrieval (MIR) constitutes a very active multidisciplinary research area that is being transformed into a cross-cutting field. Digital libraries, bio-computing & medical science, the Internet, streaming video, databases, cultural heritage collections, and peer-2-peer networks have created a worldwide need for new paradigms and techniques on how to structure and search multimedia collections. For this reason, MIR systems are one of the most promising fields in the area of information management.

Traditional IR systems only deal with textual and unstructured data. As a result, they are unable to support different kinds of media that are considered typical of a MIR system. For example, multimedia documents are more difficult to index than textual documents, and while metadata is crucial for MIR systems, traditional IR systems do not have such requirement.

The most important feature of a MIR system is the variety of data it must be able to support. Multimedia systems must have the capability to store, retrieve, transport, and present data with very heterogeneous characteristics such as text, images (both still and moving), graphs, and sound. For this reason, the development of a multimedia system is considerably more complex than a traditional information system. Indeed, the latter only deals with simple data types, such as strings or integers, whereas a multimedia system must be able to support objects of very complex structure.

1.2.3 Exploratory Search

Modern IR systems are faced with new challenges. Firstly, there is the increasing complexity of the content, as described above. Secondly, the users' expectations get more exigent and ambitious. For example, it is no longer sufficient that a search engine finds a video document; users expect to get the exact position inside that video. Thirdly, the search behavior of people changes. Generally, users do not know exactly what they are looking for. The search habits of people are still surprisingly old-fashioned, reminiscent of the old days when customers of a library would walk along the shelves and browse through heaps of books.

Today, users still act in a similar way when it comes to computer-assisted searches. The expression *exploratory search* describes users who enter some keywords, randomly click on a link in the result list, inspect that document, which might be a Web page or any other type of document, and quickly decide whether this document is "good" or not. If it is not pertinent, users very frequently enter new keywords without checking other links from the result list [FDD+99, Blo01, HS00].

This paradigm shift of user search behavior requires new knowledge representation and retrieval technologies as well as new retrieval strategies. Some promising technologies have emerged from the Semantic Web movement, e.g., ontologies, Linked Data, Resource Description Framework (RDF), and Web

Ontology Language (OWL). Such technologies allow to describe multimedia documents with machine-readable metadata in a way that applications such as search engines are able to process and integrate into their processes.

For a search engine, the meaning of content is more important than the content itself. Also, the interpretation and the disambiguation of data get more reliable when a machine is able to "understand" the meaning of a query. Structured vocabularies and external language sources, such as dictionaries or thesauri, are used for processing natural language (NL) query terms.

Modern *expert systems* have optimized search functionalities over online and offline domain knowledge bases, turning them into reliable and helpful companions. An E-Librarian Service, which we discuss in great detail in this book, is an example of such a system.

1.3 E-Librarian Services

1.3.1 Overview

An E-Librarian Service is a computer system that is able to retrieve multimedia resources from a knowledge base more efficiently than if it were to browse through an index or perform a simple keyword search. The premise is that more pertinent results would be retrieved if the *semantic search engine* "understood" the sense of the user query and was able to reason over the data. The returned results would then be logical consequences of an inference rather than of keyword matching.

An E-Librarian Service allows users to enter complete questions in NL and retrieves only few, but semantically pertinent answers. It is able to generalize or specialize the query in order to find the most appropriate document(s) in its multimedia knowledge base. This is particularly interesting if the system finds out that there is no document that will deliver a complete answer. In this case, the E-Librarian Service identifies and retrieves documents that are semantically closest to the query, based on the premise that people always expect an answer, even if it is not a perfect one.

An E-Librarian Service does not return the answer to the user question, but it retrieves the most pertinent document(s) in which the user can find the answer to his question. For example, let us suppose that the user asked the following question:

What are the tasks of a protocol? (1.1)

Just like a human librarian, the E-Librarian Service would then deliver a document in which the user finds the answer to his question. This librarian approach to solving complex retrieval problems will be described in more detail in section 1.3.4.

An E-Librarian Service is an ontology-driven expert system about a given domain, e.g., computer history, fractions in mathematics, or networks in computer science. It relies on specialized and hierarchically organized knowledge bases and specific reasoning services. The documents in the knowledge base are described by metadata that is encoded in a knowledge representation formalism, like the *Web Ontology Language* (OWL).

1.3.2 Early Question-Answering Systems

Question-answering (QA) is a type of IR. Given a collection of documents, such as the WWW or a local database, the system should be able to retrieve answers to questions asked in NL. QA is regarded as requiring more complex natural language processing (NLP) techniques than other types of IR, such as document retrieval, and it is sometimes considered the next step beyond search engines.

Some of the early artificial intelligence systems developed in the 1960s, such as *Baseball* and *Lunar*, were QA systems. *Baseball* answered questions about the US baseball league over a period of one year. *Lunar* answered questions about the geological analysis of rocks returned by the Apollo moon missions. It was demonstrated at a lunar science convention in 1971 and it was able to answer 90% of the questions in its domain asked by people unfamiliar with the system.

The 1970s and 1980s saw the development of comprehensive theories in computational linguistics which led to the development of ambitious projects in text comprehension and QA. One example of such a system was the *Unix Consultant* [WAC84], a system that answered questions pertaining to the Unix operating system. The system had a comprehensive hand-crafted knowledge base of its domain and it aimed at phrasing the answer to accommodate various types of users. The system developed in the Unix Consultant project never went past the stage of simple demonstrations, but it helped the development of theories on computational linguistics and reasoning.

1.3.3 Natural Language Interface

The interaction between humans and computers is still surprisingly complicated. Finding information in a knowledge base often means browsing through an index or formulating computer-readable queries. In both cases, the users must adapt themselves to the machine by entering precise and machine-readable instructions.

However, most people are not search experts and therefore have difficulties when it comes to formulating their queries in a machine-optimized way, e.g., by combining search terms with Boolean operators. Furthermore, they might not use the right domain expressions, which but adds to the complexity of the problem.

An NL interface simplifies the human-machine interaction. An E-Librarian Service allows users to freely formulate questions in NL, which allows them to focus on *what* they want, rather than to worry about *how* and *where* to obtain the answer.

In order to create an "intelligent" search mechanism, the user must enter a query which contains enough semantics, so that the E-Librarian Service "understands" the sense of the question and is able to logically infer over the knowledge base. A complete question in NL contains more semantics than just keywords. An E-Librarian Service uses linguistic information within the user question and the given context from the domain ontology in order to "understand" the sense of the sentence and to translate it into a logical form.

1.3.4 No Library without a Librarian

Let us suppose that Paul wants to find out some information about the invention of the transistor. He goes to a library and asks the librarian: "I want to know who invented the transistor." The librarian perfectly understands Paul's question and knows where to find the right book. He also understands that Paul does not want all the available books in the library that explain how a transistor works, or those which illustrate in detail the lives of its inventor(s). It is evident for the librarian that Paul only wants one pertinent document in which he can find the answer to his question. This illustration leads to the following statements:

For the client:

- Paul formulates his question in NL.
- Paul has no knowledge about the internal organization of the books in the library.
- Paul does not know what he is looking for in particular, e.g., he does not give the librarian a precise book title.

For the librarian:

- He is able to understand the client's question (both language and meaning).
- He does not know the answer to the client's question.
- He controls the internal organization of the library.
- From all the existing books in the library, he finds the one(s) that best fit(s) the needs of the client.

It is obvious that the larger the library is, the more documents will be potentially pertinent, especially if general questions are asked. If Paul wants to be sure that he will only get a very short list of relevant books, then he should formulate a more precise question or go to a specialized library. There, the potential amount of documents is far smaller, but the chance of finding pertinent results is higher. Visiting specialized libraries also reduces the risk

of ambiguity. If Paul asked for a book about "dragons" in a general library, the librarian would have the choice between a mythical creature and a musket. However, if Paul were in a library dedicated to weaponry, the context would be clear.

1.3.5 Characteristics of an E-Librarian Service

An E-Librarian Service is a computer-based expert system that offers the same services as a real librarian. The core part of an E-Librarian Service is an MIR module that performs a semantic search over the knowledge base. It retrieves only few but semantically pertinent documents as an answer to the users' questions. One should not confuse it with a software to manage a library or with a search engine over a catalogue.

Unlike classical search engines or QA systems, an E-Librarian Service does not deliver the answer to the user's question, but it is able to find and retrieve the most pertinent document(s), in which the user will then find the answer to his question.

Let us stretch out the difference between both approaches. The first category of retrieval systems seeks to provide concise and succinct answers to NL questions. The aim of such search engines or QA systems is to perform fine-grained, targeted IR. For example, let us consider the following question:

Who invented the transistor? (1.2)

A QA system should return a precise answer, i.e., the names of the inventors of the transistor: William Shockley, John Bardeen, and Walter Houser Brattain.

The second category of retrieval systems, to which E-Librarian Services belong, implement strategies for extracting documents or sub-parts of documents that contain the answer to the query. Such retrieval strategies are generally known as *passage retrieval*. They have been used in classical IR systems over textual documents and have proven effective when documents are of a considerable length or when there are topic changes within a document.

An E-Librarian Service is a reliable and easy-to-use expert system that allows users to find pertinent resources in a multimedia repository very quickly. Its concept adheres to the stream of the Semantic Web philosophy and joins the efforts to standardize and link reusable multimedia content, ontologies, and technologies.

An E-Librarian Service improves domain ontology search engines; fewer, yet more pertinent results are returned, as will be demonstrated when we present different applications as "best practices".

An E-Librarian Service can easily be used in other areas, such as online helpdesks or travel planners. Clients requiring assistance, e.g., with their Internet connection, or with itineraries, could contact a "virtual online help desk" and express questions in NL. The E-Librarian Service will then "understand" the gist of the customers' questions and suggest short, but pertinent answers.

Here is a summary of the characteristics of an E-Librarian Service:

- It has a huge amount of stored knowledge in multimedia form.
- It controls the internal organization of its knowledge base.
- It "understands" the gist of the user questions.
- A query is expressed in NL.
- Given a query, it finds pertinent documents in its knowledge base.
- It is able to visualize the pertinence of the delivered documents, i.e., ranking of the results according to their semantic relatedness to the query.
- It is simply accessible without complicated software or hardware requirements.
- It is simple to use, i.e., the interaction takes place in a human way by means of verbal communication.

1.4 Overview and Organization of the Book

This book focuses on the design, implementation, and testing of a novel approach to retrieval systems by regrouping the most appropriate theories and technologies from different research domains. It also described the required technologies, strategies and technical details on how to develop an E-Librarian Service.

This book is structured as follows:

- Part 1 presents the technologies required to build an E-Librarian Service. In chapter 2, the Semantic Web and its underlying technologies (ontologies, XML, RDF, and OWL) are explained. Chapter 3 focuses on Description Logics as formal knowledge representation language. An introduction to NLP and to IR is given in chapters 4 and 5..
- Part 2 of this book is dedicated to the design and implementation of an E-Librarian Service. Chapter 6 focuses on the ontological approach and the creation of metadata, while chapter 7 describes how to develop the NLP module. The design of the MIR module, i.e., the semantic search engine is explained in chapter 8. Implementation details are described in chapter 9.
- Part 3 of this book illustrates three best practices of E-Librarian Services that were tested in real life scenarios. These are the *Computer History expert system* (CHESt), *Mathematics expert system* (MatES), and the Lecture Butler's E-Librarian Service.
- Part 4 concludes this book with an appendix of different resources that are not mentioned in the main content.

http://www.springer.com/978-3-642-17742-2

E-Librarian Service
User-Friendly Semantic Search in Digital Libraries
Linckels, S.; Meinel, C.
2011, XVI, 212 p., Hardcover
ISBN: 978-3-642-17742-2