Topic driven access to scientific handbooks

Caracciolo, C.

CHAPTER 1

Introduction

Mary Ann has to write an essay on Gödel's theorem. Prof. Dickenson needs to check a bibliographic reference. Robert came across the expression 'categorial grammar' and wants to find out what it means. Dr. Schultz is updating the bibliography for an old paper of his. Susan wants to find out what the prominent centers of research in the area of computational linguistics are... What is common to all these scenarios is a question mark: the fact that they can be turned in (at least) one question, either vague or very sharp, simple or complex, ambitious or minimalist. For the rest, each of the above situations involve different people, with different backgrounds, aims, expectations about the type of answer to get, and in what time. Also, their ability to select and interpret answers, and their notions of relevance and satisfaction differ.

What is usually called an information need, what pushes a person to look for information, is the result of all these factors. When using a search engine, or OPAC system, the user has to type in a *query*, a phrase that instantiates the user's information need. Queries are not complete sentences, but phrases containing the salient terms the person is interested in. Formulating a query implies knowing what one is looking for, and formulating a good query implies having an idea of what documents satisfying the need should look like. Therefore, the query is a fundamental component of an *information searching* process, but information search is included in the more general process of *information seeking*, which is one aspect of a person's *information behaviour* (Figure 1.1). Many models have been proposed [Dervin, 1983, Ingwersen, 1984, Marchionini, 1995, Wilson, 1994] to describe what happens when a user seeks information, starting from the time the information need is perceived, and going through all passages necessary to achieve the goal (or to give up on it). Some of them (e.g., [Wilson, 1997]) also include a temporal dimension in their model for information behavior: there, the "active search" is the last component of the (iterative) process of information seek-

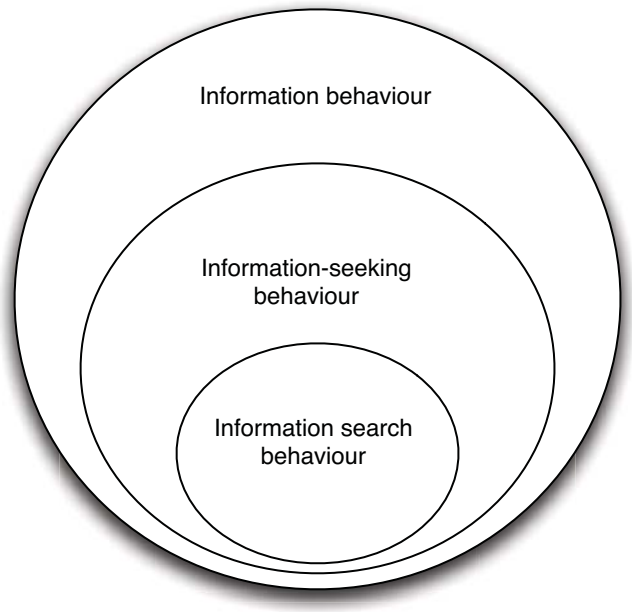

Figure 1.1: Nested model of information behavior, according to Wilson [1999].

ing. The point to make here is that the information seeking behavior has search as an important, but not unique, component. The intertwined browsing, exploring, reading activities all play a role in satisfying an information need, and should be supported in a good electronic publishing environment. In such an environment, the reader will be naturally brought "to" and "inside" the text.

In this thesis, the type of text we are interested in is scientific handbooks, a type of text that has so far received relatively little attention from the library science or information retrieval community. We address the issue of providing topic driven access to scientific handbooks, and we do so by adopting a broad perspective. We present a model that includes both browsing and searching as fundamental components, and that builds on ideas and approaches developed in the areas of library science and information retrieval.

1.1 Problem Statement

The tradition of the scientific handbook as a concise, accessible source of validated information emerged in the late nineteenth century when the factual burden of scientific and medical subjects began to overwhelm students. Nowadays, handbooks are "fat" volumes that typically contain long chapters, written by many authors and of several dozen pages each, often without a standardized structure, with the aim of providing a comprehensive overview of a scientific discipline. Now, in the electronic age, access

to these (electronic) scientific handbooks themselves has become an issue requiring attention.

Users of scientific handbooks may have a specific information need, which leads them to read up on the specific topic and even on a specific aspect of it. If this is the case, such readers want to avoid having to read or scroll dozens of pages, they rather need a way to "jump" to specific excerpts of the handbook covering the topic they are looking for. This is what we call *focused access* to the text: when the reader is brought directly inside the text, to the specific excerpt where she can find the information she is looking for. In other cases, users of handbooks may have vaguer information needs, more related to the need or desire to get a more general, higher level picture of the domain.[1] In the former case, the user is likely to be able to produce a query, while in the latter case this ability may be hampered by poor knowledge of the domain.

These considerations naturally lead one to think of an environment where the user is not only able to type in queries, but can also take advantage of an organization of the material that can meet her vague information needs. In reference works, such as dictionaries and encyclopedias, subjects are usually arranged in alphabetical order so that they may be located quickly and easily. In books and handbooks, the subject is organized into chapters, and usually an index is provided to serve as the direct guide to the many topics treated in it, or to locate the smaller subdivisions of the larger subjects. The traditional back-of-the-book index may also be organized in several disjoint parts, such subject index, author index, name index, etc. Another important element in the arrangement of material in a reference work is the *cross-reference* that will refer the reader to additional related information. From the back-of-the-book index, then, comes the inspiration to organize the subject in a way that is informative for users with vague information needs and a limited background in the area. In this setting, it is natural to think of a high-level *map* of a domain, a searchable map containing topics and relations between them, as well as appropriate locations inside the text. Then the user can "land on" the map, either by search or by navigation, and then zoom in on the topics that best fit her information need. Therefore, we propose a map of the domain, that beyond containing references to the appropriate locations in the text, also includes relations between the elements in it — that we call *topics*.[2] So, the fundamental question is: can we gain anything from providing a map enriched with links to access scientific handbooks?

To make matters more concrete and tangible in the thesis, we work with a specific domain and handbook: the domain is the interface of logic and linguistics, and the handbook that we use as our test case is the *Handbook of Logic and Language* [van

[1] We do not consider here the case of consecutive reading: this does not mean that it is not an important type of reading or, worse, that it has been put out of fashion by the advent of the "electronic age." On the contrary, its role for educational purposes cannot be overestimated, but in our work we simply look at a different usage of scientific handbooks.

[2] Note that *topic maps* [TOPICMAPS, 2006] is an ISO standard for describing knowledge structures and associating them with information resources. It is essentially a format for data representation and can only express two types of *associations* between topics: *class-instance* and *superclass-subclass*. In this sense the LoLaLi map of topics is very different from a topic map.

Benthem and ter Meulen, 1997]. In this domain we identify important topics and relationships between those topics; these are used to build a browsable map, which is then presented to the reader as an interface to the handbook.

Given this approach, that we explain in more detail in Chapter 2, the research questions we address in this thesis are the following:

---

1. RESEARCH QUESTION. What requirements should we impose on a map that is to be used for human browsing and as a skeleton to provide focused access to the text of a scientific handbook?

In the case of the *Handbook of Logic and Language*, Research Question 1 comes with an important constraint: the process of populating the browsable map should be as much as possible a collaborative and bottom-up process. There are several reasons for this constraint, some principled, some pragmatic. The interface of logic and language is evolving rapidly, and is highly interdisciplinary. Moreover, unlike, say, medicine or law, there are no resources (or standard bodies) to support imposing a standard in a top-down manner: (expert) colleagues around the world, and from around the interdisciplinary area covered by the *Handbook of Logic and Language*, will be the ones populating the browsable map with topics and relations between them.

---

2. RESEARCH QUESTION. How do we present the map to readers of a handbook in such a way that we ensure broad coverage of the domain (with detailed information per topic), while also making sure that users do not get lost?

The answer to Research Question 2 will be a user interface that is able to accommodate detailed pieces of information about topics in the map, while at the same time giving the user ease of navigation and a good sense of the "general picture." The interface should also be able to equally support searching and browsing in order to allow users to satisfy a broad variety of information needs, from the more specific (search) to the more vague (browsing), and user background.

Assuming that we are able to come up with a satisfactory answer to Research Question 2, we need to connect the browsable map to the documents of which it is meant to provide a map. The next question, then, is:

---

3. RESEARCH QUESTION. What are suitable targets in the handbook to establish focused, topic driven links from topics in our browsable map?

Notice that in Research Question 3 we ask for *focused* links from topics in the browsable map into the handbook, and not just for links from a topic to an entire chapter of dozens of pages. A focused link, then, should be an excerpt of the document, *readable* despite its separation from the whole of the document, and relevant to the topic it talks about. Assuming there are different ways to identify excerpts that could provide the basis for focused links, we should find out which is the most appropriate. We phrase this desideratum as follows:

4. RESEARCH QUESTION. What is the most suitable type of candidate link to be connected to the map?

An obvious key issue along the way will be *evaluation*: how do we assess our proposals? At different stages of our work, different types of evaluation are appropriate. When we introduce our proposal for a browsable map that generalizes the concept of the back-of-the-book index to the setting of electronic scientific handbooks, we compare the results we obtained with the requirements we set ourselves, but also with the internal organization of the map. Later, we perform user studies to assess the effectiveness of our proposed visualization method for exploring the map. A third type of evaluation is conducted when we link the topic in our browsable map to candidate targets in the handbook; this is the type of evaluation that one encounters in information retrieval and applied language technology, where one develops "gold standard" corpora of ideal outcomes and measures the performance of algorithms against this yardstick. Since no standard test sets exist for the issues that we are tackling, we develop our own.

Throughout the thesis, we will have an additional "meta-concern" on our mind: assuming that our proposed model for accessing electronic scientific handbooks is an effective one, how are the roles of authors and editors affected?

5. RESEARCH QUESTION. Can the scenario proposed in this thesis be adopted as a basis for the production of new handbooks and, if so, would this imply a change in the roles of authors and editors of handbooks? Will they be expected to populate the envisage browsable maps? Will they have to write differently, knowing that a map will be linked to their texts?

1.2 Organization of this Thesis

In this thesis we proceed in a step-by-step manner by providing answers to each of the research questions listed above. Figure 1.2 provides a graphical representation of the contents of the thesis. In Chapter 2 we describe our vision of focused access to scientific handbooks: we propose the idea of a browsable map displaying important terms and their interrelations as a suitable interface to electronic scientific handbooks. We also provide general background on the issue of searching for and inside books, and on structures used for indexing and classification purposes.

Chapter 3 addresses Research Question 1. There, we instantiate the proposed model to our case study: the domain of logic and language. We introduce our map, that we call the *Logic and Language Links* map (for short, the LoLaLi map), the way it was built and its constituents, including the hierarchical and non-hierarchical relations between topics. In the same chapter, we also report on the management of the map.

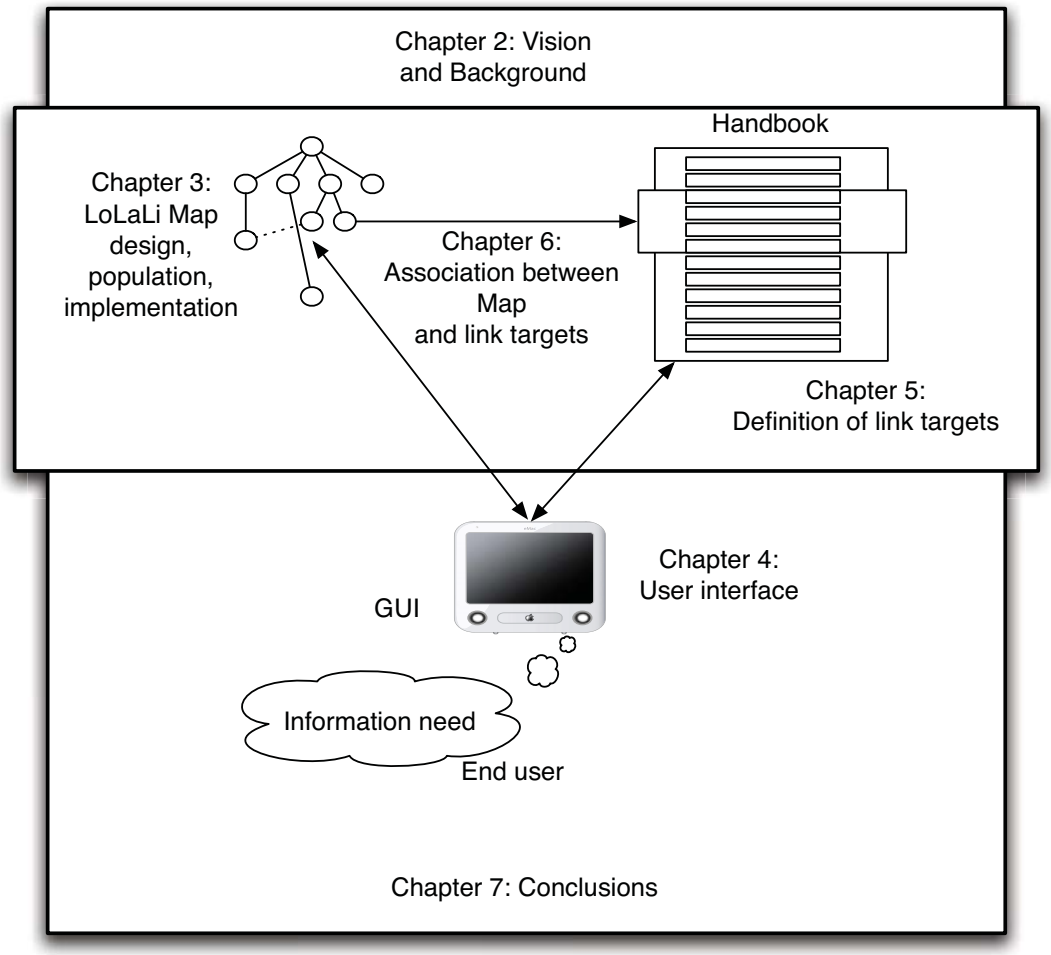

Figure 1.2: A graphical representation of the contents of this thesis.

Then, in Chapter 4 we address Research Question 2. We consider the problem of presenting to end users the browsable map we developed. We first review well-known visualization techniques for semantic structures like ours, then we describe the interface we developed in order to allow end users to navigate and search the LoLaLi map. Finally, we report on the user studies we performed with two groups of students.

In Chapter 5 we concentrate on the selection of text excerpts to be used as candidate link targets for the LoLaLi browsable map, thus addressing Research Question 3. We take into consideration two types of automatic text segmentation techniques, structurally and semantically oriented, and we compare them against a manual annotation of the handbook.

Given the collection of text segments resulting from the experiments described in the previous chapter, in Chapter **??** we address Research Question 4 and look at the matching between these segments and the topics in the map. We discuss the evaluation

of that task and propose measures for this purpose. Finally, in Chapter 7 we draw our conclusions by summarizing our answers to Research Questions 1–4, and address Research Question 5. In Chapter 7 we also highlight our plans about future work.

It should be noted that this thesis does not contain a monolithic "Related Work" chapter. We decided against this: as we build on insights from a broad range of disciplines—including library science, user interface design, human computer interaction, information retrieval, and knowledge representation—such a chapter would have been a maze of disconnected discussions. Instead, we review related work and link our own contributions to the literature along the way.

1.3 Scope of this Thesis

In order to address our research questions we build on insights and methods from library science, user interface design, human computer interaction, information retrieval, and knowledge representation. Our work aims at exploring a possible way to provide focused, topic-driven access to scientific handbooks, in this respect it cannot be easily classified under any of these labels. Moreover, we do not aim at providing an end-to-end system. For these reasons, the following issues are, unfortunately, beyond the scope of this thesis.

Production. Many ingredients are necessary to have an end-to-end systems up and running, including a session manager, text encoding, large data management systems, efficient server-client interaction, integration of search over several sources and visualization and rendering. Although necessary, these components are deliberately not included in the framework presented in this thesis.

Publishing. We do not deal with issues such as copyright and intellectual property or business and publishing models for online publication of reference works. We believe these issues are central and need to be addressed so as to promote a wide and democratic circulation of knowledge and thoughts, while guaranteeing authors the possibility of continuing to create intellectual work.

Handbook authoring environments. Although in the concluding chapter we dedicate some space to future electronic environments for authoring handbooks, this thesis mainly deals with the conversion of existing handbooks in order to provide topic access to them.

E-learning. Any handbook is both a reference tool and a learning support tool, and in this thesis we concentrate on accessing electronic handbooks in their function as reference tool. Issues related to the learning function of an electronic handbook, such as cognitive aspects of online reading and online learning, are therefore not treated here.

1.4 Origins of the Material

The material in this thesis is largely based on a number of publications. Chapter 2 includes material published in:

- C. Caracciolo and M. de Rijke (2002). Structured Access to Scientific Information. In *Proc. of Global WordNet Conference*, Mysore (India).

The description of the LoLaLi map given in Chapter 3 is based on previous work published in:

- C. Caracciolo (2003). Towards Modular Access to Electronic Handbooks. *Journal of Digital Information (JODI)*, 3(4), no. 157. URL: http://journals.tdl.org/jodi/article/view/jodi-104/84.

- C. Caracciolo (2006). Implementing an Ontology for Logic and Linguistics. *Literary and Linguistic Computing*, 21:29–39.

The prototype used for the user studies described in Chapter 4 used tools developed by W.R. van Hage and described in:

- W.R. van Hage (2004). Living on the Edge. Master thesis, University of Amsterdam, 2004.

The results presented in Chapter 5 have been published in:

- C. Caracciolo, M. de Rijke, and J. Kircz (2002). Towards Scientific Information Disclosure Through Concept Hierarchies. In J. A. Carvalho, A. Huebler, A. Baptista, editors, *Proc. of the 6th International ICCC/IFIP Conference on Electronic Publishing* (ELPUB 2002), Karlovy Vary (Czech Republic).

- C. Caracciolo, W. van Hage, and M. de Rijke (2004). Towards Topic Driven Access to Full Text Documents. In R. Heery and L. Lyon editors, *Proc. of European Conference on Digital Library (ECDL 2004)*. Bath (UK). Springer Verlag.

The results presented in Chapter **??** have been published in:

- C. Caracciolo and M. de Rijke (2006). Generating and Retrieving Text Segments for Focused Access to Scientific Documents. In M. Lalmas, A. MacFarlane, S. Ruger, A. Trombos, T. Tsikrika, A. Yavlinsky, editors, *Proc. of European Conference on Information Retrieval (ECIR 2006)*. London (UK). Springer Verlag.