Supporting Polyrepresentation and Information Seeking Strategies

Thomas Beckers
Department of Computer Science and Cognitive Science
Universität Duisburg-Essen
47048 Duisburg, Germany
<http://www.is.inf.uni-due.de>
tbeckers@is.inf.uni-due.de

Abstract

This paper introduces the basic concepts and notions of a new framework for interactive information retrieval. Based on examples for a real-life collection of book data we show why current systems are not sufficient. It is necessary to support both polyrepresentation of information objects and multiple information seeking strategies in order to cope with the shortcomings of most current retrieval systems. A search operator concept is introduced which controls the process of retrieval. We provide real-life examples of how information seeking strategies can be supported. Furthermore, we show why interaction should play a more important role than it does today. Finally, we give an outlook about our future plans and upcoming research challenges.

Keywords: interactive information retrieval, polyrepresentation, information seeking strategies

1. INTRODUCTION AND MOTIVATION

Most existing information retrieval systems support a very limited view on documents. Usually, only the content of a document is regarded, which is made searchable by using a single representation. Furthermore, in the majority of cases there is just limited support for interaction between the user and the system.

Based on an analysis of information retrieval as an information seeking activity, Belkin [1] suggests that interactions with documents should play an important role during the retrieval process. He defines a set of information seeking strategies (ISSs) and points out that different ISSs require different interactions and search behaviours. Thus, traditional user interfaces and systems that are solely based on simple representations of texts and assume well-defined queries cannot satisfy all possible information needs of the user.

In this paper, we propose a concept that deals with the aforementioned challenges and that aims at improving the retrieval process by incorporating different aspects of documents and supporting a variety of information seeking strategies.

As an application example for demonstrating our ideas, we regard Amazon's¹ book catalog in combination with LibraryThing's² book metadata. At the moment, it is not possible to search in Amazon e.g. for the content of books or front covers and neither for data about the dimensions nor for the number of pages of books. Although searching in customer reviews is possible in principle, it is very restricted and deeply hidden inside Amazon³. LibraryThing provides even less possibilities for book search. In both systems the interaction is restricted to defining the search query and sorting the result list according to certain attributes like e.g. publication date or price. Both search interfaces can be considered as exemplary for many current real-world systems.

¹<http://www.amazon.com>

²LibraryThing's is a social book cataloging web application which allows its user to store and share their personal library catalogs, <http://www.librarything.com>

³There is a text field at the review detail page of a book. If the checkbox is disabled a search can be performed on all reviews. All statements concerning Amazon and LibraryThing are as of April 2009.

For demonstrating our ideas, we built a new test collection by crawling the meta-data of approx. 2.7 million books from Amazon as well as from LibraryThing and merged them into a coherent structure. The data from Amazon includes the creators (e. g. author, editor, illustrator), title, publisher, dimensions (height, width, length, weight), classifications (reading level, subjects, browse nodes), thumbnails of the cover, similar products, editorial and customer reviews. LibraryThing supplies user generated data about blurbers⁴, dedications, epigraphs, first words, last words, quotations, series, awards, people, places and tags.

The aim of our proposed framework is to expand classic retrieval systems by allowing the user to interactively use rich representations of documents for retrieval while as many ISSs as possible are supported.

2. RELATED WORK

Ingwersen proposed the concept of *polyrepresentation* as a general framework for interactive information retrieval [2]. All participating cognitive structures are of potential value on both the system side and the user side. To support different cognitive structures as many and as different representations of information objects (documents) and information needs as possible should be used for retrieval. This principle of intentional redundancy is called polyrepresentation. Ingwersen's work supports our hypothesis that single, simple representations of documents are not sufficient to allow effective information retrieval.

Belkin et al. [3] propose four facets for classifying ISSs. In this paper, we focus on the three facets *method*, *mode* and *goal* of seeking. The method can be either *searching* or *scanning*, whereas searching refers to a targeted and focused search while scanning is the mostly sequential examination of a result list. The mode can either be *specification* if the user is able to express his information need or *recognition* if he is not. The goal can either be *learning* about e. g. documents or *selection* of documents. A first approach for a system supporting multiple ISSs is described by Yuan and Belkin [4], which focuses on the *method* facet.

In contrast to the traditional understanding of computation the paradigm of *interactive computation* [5] does not reduce every task to a simple function but regards interaction as an important part of solving a task. Wegner [6] and others claim and circumstantiate that interaction can be more powerful than classic algorithms. Brought forward to the domain of information retrieval, interaction should play a bigger role. Currently, most research still focuses on the query formulation – result computation cycle and aims at optimising the latter. Instead, we should allow for richer interaction possibilities, making interactive retrieval more flexible such that it can adapt to the different ISSs.

3. POLYREPRESENTATION, SEARCH OPERATORS AND INTERACTION

We think that there are three important concepts which are essential for effective interactive information retrieval supporting different ISSs: polyrepresentation, search operators and interaction, which we discuss in the following.

3.1. Polyrepresentation

Our notion of polyrepresentation is broader than Ingwersen's, by comprising as many aspects of information objects as possible. Fig. 1 shows an example of our interpretation of this concept. Every facet of a *document* can be modelled by a so-called *aspect*. For each aspect, there may be various *representations*, which form the reference points for searches. In our opinion, this broad notion of polyrepresentation is more adequate for the type of information objects we are dealing with today.

The Lacostir project [7] suggests three aspects, namely layout, content and structure. The content is what classic information retrieval is about. The structure aspect comprises the logical structure of documents as it is common for database-oriented XML retrieval like e. g. in XQuery without full-text search; classical examples of this aspect are document attributes like e. g. author or publication year. The layout aspect is concerned with the displaying of documents on mediums. These aspects can serve

⁴A blurb is someone who writes a very short – mostly positive – review for the back cover of books.

FIGURE 1: Our notion of polyrepresentation of information objects described by a sample book of our collection from Amazon/LibraryThing

as a generic polyrepresentation that fit to nearly all possible documents. Depending on the concrete application these aspects can be redefined or new aspects and representations can be added.

In Fig. 1, a possible definition of aspects for our Amazon/LibraryThing book collection is depicted. The content aspect incorporates representations of the actual content of a book, such as tags or editorial reviews. Representations of the structure (e.g. author, publisher) are part of the structure aspect. Several representations of thumbnail images of a book cover belong to the cover aspect, like e.g. color histograms or spatial color distributions. Since Amazon allows its users to write reviews about books, the aspect reviews contains all representations of reviews such as the actual content or the rating ranging from one up to five stars.

3.2. Search Operators

Our *SOPV* model for interactive retrieval consists of four steps, namely *selection*, *organisation*, *projection* and *visualisation*, hence its name. We call operations in each of these steps *search operators*. These operators model the interactive process of information retrieval. The steps of this model correspond to the reference model for information visualisation [8] proposed by Shneiderman. Due to polyrepresentation of documents with respect to their aspects, different operators are essential for different representations of documents. Following, the four steps and some possible operators are described in some more detail.

Selection A user formulates selection conditions that pick possible relevant documents from one or more document collections. This step includes the choice of search queries, retrieval models and document collections. It is possible to search with respect to the various aspects but of course only on available representations (e.g. searching at Amazon for a book showing a high-speed train on the front cover is not possible).

Organisation The selected documents can be organised in various ways. Possible forms of organisation are the traditional list sorted by retrieval status value as it is used by most current retrieval systems or clustering the results by similarity of certain representations. The choice of the organisation also depends on the choice of aspects. Sorting documents according to the content is not possible while sorting after certain structure aspects (e.g. publication date) is reasonable. Consider a thumbnail of a book cover and tags describing the book's content. Performing clustering on those representations would require a variety of appropriate clustering operators [9]. So, the set of possible operations depends on the types of the representations that are used for organisation. Summing up, a user can e.g.

- organise the selected documents as list or as a 2/3-dimensional table or space, sorted by certain attributes or
- perform clustering with regard to some representations or attributes thereof.

Projection The user may be only interested in certain attributes of representations, e.g. the title and the authors of the structure aspect. This filtering is applicable on structure attributes while other types of representations should allow for different projections. A possible projection for content representations (e.g. the content of reviews) are query-based summaries which project the content to a short summary related to the search query (see e.g. [10]). As described in the two examples above the choice of projection operations depends on the type of the representation.

Visualisation Finally, the selected, organised and projected representations are visualised. This visualisation step is required given that there are countless possible visualisations for a single visual structure which is the result of the steps performed before.

This model can be described best by means of an example based on our book collection (see fig. 2). A user needs information about concurrency in Java. He heard about a good book about concurrency in Java with an express train on the front cover some months ago. He can't remember the name of the author but he would be able to recognise him if he reads his name. He *selects* books that match his query `java concurrency`. Then, he *organises* the resulting books as a list ordered by the retrieval status value whereas he *projects* only on attributes he's interested in: authors, title and a thumbnail of the front cover. Finally, the user chooses a *visualisation* that is similar to that of Amazon. The first two books seem to be relevant. He's now able to recall the name of the author, namely Brian Goetz. So, the first book is the one he is searching for.

FIGURE 2: Example of the SOPV model based on our book collection

3.3. Interaction

The SOPV model offers many possibilities for interactions: the choice and configuration of the three search operators as well as interaction with the visualisation. Classic interaction techniques include panning & zooming, focus + context and highlighting (e.g. of terms contained in the search query). Also, more advanced interaction techniques are possible, such as query by example by allowing the

user to specify search and projection operations through exemplary marking of certain attributes of representations: Given an example document, the user could edit some of its attributes and specify them as search conditions [11]; furthermore, by highlighting certain parts of the entry, the user could indicate that only these parts should be shown for each result item.

4. SUPPORTING INFORMATION SEEKING STRATEGIES

The various ISSs can be supported by different combinations and configurations of search operators and interaction. There are many possible types of different ISSs. Following, some examples of ISSs referring to the aspects are outlined.

- (i) I am looking for a cookbook about the Chinese cuisine that has good reviews. Thus, I start a search for books about chinese cuisine. Only books with a rating greater than 4 stars should be shown. The result list should be sorted by the average rating (searching, specification, select / content and review aspects)
- (ii) I need a book as a gift for my girlfriend. I only know for sure that her favourite books are novels about a police inspector in a Scandinavian country. So, I want to cluster novels by their content. The most important terms should also be shown in order to recognise relevant books. (scanning, recognition, select / content aspect)
- (iii) A friend of mine has a guidebook about New Zealand that I want to use during my next holidays. I want to know if it covers all places which I want to visit. Therefore I search for a guidebook via authors and title to learn if a summary of this book's content contains terms like e.g. milford sound (searching, specification, learning / content aspect)

For supporting (i) the user may want to project only on attributes, like e.g. the author and the title summary, that are well suited for identifying the relevant books while supporting (ii) the user may decide to use additional projections, e.g. a query-based summary of the book's content, and a different organisation, e.g. a list sorted by publication date. If he can remember the front cover (cover aspect), he would probably apply an organisation operator which clusters books by similarity of their front covers.

Amazon only offers limited support for ISSs. Searching as method of seeking is only possible based on the structure aspect of books. One can organise and project according to e.g. the publication date while searching in reviews or cover images is not possible. Amazon does not offer operations on the cover and review aspect because there are no representations for these aspects (i) to create appropriate surrogates of documents that allow searching. While searching is not well supported on these aspects, scanning on the cover aspect is possible since there are thumbnails of them allowing scanning. However, scanning for content (ii) or review aspects is impossible because there are no operations to create adequate document surrogates. Learning as goal of seeking (iii) as well as recognition as mode of seeking (ii) is not directly supported at all.

Overall, we see that it is technically possible to create polyrepresentations covering all document aspects, which can be used for defining appropriate search operators. Thus, we want to develop an interactive framework, which in turn forms the basis for supporting ISSs.

5. USER INTERFACES FOR INTERACTIVE RETRIEVAL

Due to the variety of polyrepresentation and ISSs different or flexible configurable user interfaces are required. For instance, an ISS that needs to visualise clusters of documents needed for scanning or recognition must be treated with a different user interface than an ISS that relies on the retrieval of known items. However, the interface of most search engines (including Amazon's) are of rather static nature. The design of the optimal user interface is subject of our future research. One main challenge is to find a good balance between flexibility and complexity. A highly flexible user interface supporting many ISSs and polyrepresentation would allow maximum search and interaction possibilities but the increased complexity may confuse or overwhelm in particular unexperienced users and at the worst even experienced users like e.g. librarians.

As described above, the underlying principles of classic user interfaces for retrieval are not satisfactory, thus we aim at developing a concept for flexible yet easily usable user interfaces that support our concepts. We want to find out how an optimal user interface for interactive retrieval that supports ISSs, polyrepresentation and interaction should look like.

6. CONCLUSION AND OUTLOOK

In this paper we have outlined a new approach for interactive information retrieval. The concepts of rich and diverse polyrepresentations of documents as well as a model based on selection, organisation, projection and visualisation to support various ISSs were provided. We have laid out the basic notions that are required for effective interactive information retrieval. We have illustrated our concepts using a collection of book data.

Future research questions include whether our framework is actually adequate for effective interactive retrieval and if the SPOV model can be incorporated into a more holistic framework. It is intended to carry out a student project that aims at developing innovative user interfaces for searching in our book data. Currently, we are planning to do first implementations of our ideas based on Daffodil⁵. Therewith, we aim at performing evaluations at the iTrack of the INEX evaluation initiative ⁶ with our book data.

REFERENCES

- [1] Nicholas J. Belkin. Interaction with texts: Information retrieval as information seeking behavior. In *Information Retrieval '93.*, pages 55–66, Konstanz, 1993.
- [2] P. Ingwersen. Polyrepresentation of information needs and semantic entities, elements of a cognitive theory for information retrieval interaction. In *Proc. of the 17th Annual International ACM SIGIR Conference*, pages 101–111, 1994.
- [3] Nicholas J. Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9:1–30, November 1995.
- [4] Xiaojun Yuan and Nicholas J. Belkin. Supporting multiple information-seeking strategies in a single system framework. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference*, pages 247–254, New York, NY, USA, 2007. ACM.
- [5] Dina Q. Goldin, Scott A. Smolka, and Peter Wegner. *Interactive Computation. The New Paradigm*. Springer, 2006.
- [6] Peter Wegner. Why interaction is more powerful than algorithms. *Co. ACM*, 40(5):80–91, 1997.
- [7] Norbert Fuhr, Matthias Jordan, and Ingo Frommholz. Combining cognitive and system-oriented approaches for designing IR user interfaces. In *Proc. of the 2nd International Workshop on Adaptive Information Retrieval (AIR 2008)*, October 2008.
- [8] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors. *Readings in information visualization: using vision to think*. Morgan Kaufmann, San Francisco, CA, USA, 1999.
- [9] Kerry Rodden, Wojciech Basalaj, David Sinclair, and Kenneth Wood. Does organisation by similarity assist image browsing? pages 190–197, 2001.
- [10] Hideo Joho and Joemon M. Jose. Effectiveness of additional representations for the search result presentation on the web. *Inf. Process. Manage.*, 44(1):226–241, 2008.
- [11] K. Großjohann, N. Fuhr, D. Effing, and S. Kriewel. Query formulation and result visualization for XML retrieval. In *Proceedings ACM SIGIR 2002 Workshop on XML and IR*, 2002.

⁵<http://www.is.inf.uni-due.de/projects/daffodil/>

⁶<http://www.inex.otago.ac.nz/tracks/interactive/interactive.asp>