Moving towards adaptive search

Conference or Workshop Item

Moving towards Adaptive Search

Stephen Dignum[1], Yunhyong Kim[2], Udo Kruschwitz[1], Dawei Song[2], Maria Fasli[1], and Anne De Roeck[3]

[1]University of Essex, Colchester, UK
{sandig, udo, mfasli}@essex.ac.uk
[2]Robert Gordon University, Aberdeen, UK
{ykim1, d.song}@rgu.ac.uk
[3]Open University, Milton Keynes, UK
a.deroeck@rgu.ac.uk

Abstract. Information retrieval has become very popular over the last decade with the advent of the Web. Nevertheless, searching on the Web is very different to searching on smaller, often more structured collections such as intranets and digital libraries. Such collections are the focus of the recently started AutoAdapt project[1]. The project seeks to aid user search by providing well-structured domain knowledge to assist query modification and navigation. There are two challenges: acquiring the domain knowledge and adapting it automatically to the specific interest of the user community. At the workshop we will demonstrate an implemented prototype that serves as a starting point on the way to truly adaptive search.

1 Introduction

Document retrieval systems have been around for more than fifty years, and early systems exploited similar structures that we have in modern digital libraries, such as author name, book title, and keywords [9]. More recently we have witnessed a major shift towards search on the "Web". However, search techniques that work well on the Web do not necessarily work equally well on collections that have other characteristics, e.g. domain-specific or more structured collections, as found in intranets, digital libraries and on local Web sites. Retrieval from intranets, for example, behaves unlike Web search [14]. For instance, standard ranking functions (e.g., PageRank [3] and HITS [7]) that work well for Web collections are less effective on intranets. Furthermore, the terminology, structure, and services provided within such collections are selected to meet organisational requirements, and, consequently, a considerable amount of time is spent by users trying to learn the domain characteristics even before they are able to identify the adequate questions to be submitted to a search system. From an information systems perspective, thesauri and classification schemes should be developed and adapted to match information contained in such collections [1].

The approach that the AutoAdapt project takes is to maintain (or adapt) such structures automatically. We are looking at search as well as navigation within domain-specific document collections and our aim is to satisfy a user's information request effectively by learning from the entire user population and incorporating

[1] http://autoadaptproject.org

this learned knowledge in a constantly adapting domain model which assists a user in the search process. To support such adaptation we investigate how domain models are explored using clickthrough data that link up query modification steps and associate clicked documents with queries. This provides a context-rich environment where learning algorithms can identify new terms and relationships to add to a model, remove outdated or irrelevant terms and relationships, and modify weights as certain paths become more popular.

The workshop will be an opportunity to present a working prototype that allows system-guided search of document collections using automatically constructed domain knowledge as well as existing knowledge structures as used in digital libraries. It will also be a useful forum to discuss our current research focus, the automatic adaptation of the system using the logged user interactions.

2 Related Work

There is a wealth of related work in log analysis, interactive search and other areas, e.g., [5, 12]. Due to limited space we will only present a few findings that should serve as motivations for our own work. First of all, we know that users are reluctant to leave any explicit feedback when they search a document collection [10]. However, implicit feedback, e.g., the analysis of log records, has been shown to be good at approximating explicit feedback. For example, users often reformulate their query and such patterns can help in learning an improved ranking function [6]. The same methods have shown to improve an adaptive domain model on a local Web site [8].

We can ask, however, do users want assisted search in the first place? First of all, digital libraries are characterized by much more structured knowledge than Web sites. This makes system-guided search a natural option as evidenced by the success of Aquabrowser[2] as a tool to access digital libraries. More generally though, there is also evidence that users want support in proposing keywords but they ultimately want to stay in control about what is being submitted as a query [16]. Furthermore, despite the risk of offering irrelevant suggestions in a system-guided search system, users might prefer having them rather than not [15]. On the other hand, it has also been shown that users are more inclined to submit new queries or resubmit modified queries than to navigate from the result set in a search environment that supports search *and* navigation [11]. Perhaps the best evidence for an interactive search system is the fact that all big Web search engines have recently added more and more interactive features, e.g., Google's Wonderwheel[3].

We are in line with what Belkin calls the *challenge of all challenges* in IR at the moment, to move beyond the limited, inherently non-interactive models of IR to truly interactive systems [2]. Our aim is to go beyond static interaction patterns and move to adaptive retrieval exploiting the implicit feedback that users leave when searching and navigating a document collection. Building adaptive domain models for digital libraries and other collections is our approach to capturing and utilizing collective intelligence [13].

[2] `http://www.aquabrowser.com/`
[3] `http://www.googlewonderwheel.com`

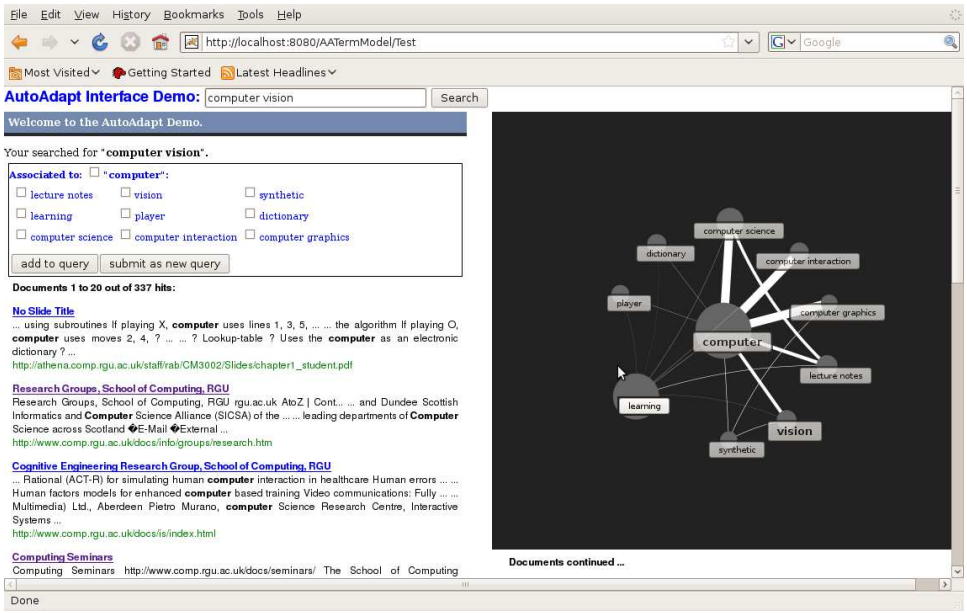

Fig. 1. Screenshot of AutoAdapt Demo System.

3 AutoAdapt Prototype

In Figure 1, we can see a screenshot of our demonstration system running on an intranet. In this particular case the domain model is automatically extracted from the document collection. The user submits a search query, this results in a number of matches (documents, book titles, etc.) being returned. Using the query terms, a segment of the domain model is displayed. The user can traverse the domain model by clicking on displayed terms. On term selection the list of suggested terms is updated. The user can then add the term to the existing query or use as a new query. The graph representations of domain models has been discussed in the literature, e.g. [4]. They are not the focus of our research but a useful tool.

The logging structure records a number of user decisions without the need for explicit feedback. What we are logging is not simply the action a user has taken (e.g., selecting a query modification, clicking a term in the domain model, selecting a match) but also recording what options a user has *not* taken but which have been available. This provides relative judgements that can be used to train classifiers [6, 8].

We have started to apply a number of techniques to turn the logged interactions into adaptive models and would like to discuss this at the workshop.

Acknowledgements

AutoAdapt is funded by EPSRC grants EP/F035357/1 and EP/F035705/1. The JIT visualisation toolkit[4] was used for the domain model visualisation.

[4] `http://blog.thejit.org/javascript-information-visualization-toolkit-jit`

References

1. M. J. Bates. The cascade of interactions in the digital library interface. *Information Processing and Management*, 38(3):381–400, 2002.
2. N. J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, 2008.
3. S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International World Wide Web Conference (WWW7)*, pages 107–117, Brisbane, 1998.
4. J. W. Buzydlowski, H. D. White, and X. Lin. Term co-occurrence analysis as an interface for digital libraries. In *Joint Conference on Digital Libraries*, pages 133–144, 2001.
5. J. Jansen, A. Spink, and I. Taksa, editors. *Handbook of Research on Web Log Analysis*. IGI, 2008.
6. T. Joachims and F. Radlinski. Search engines that learn from implicit feedback. *IEEE Computer*, 40(8):34–40, 2007.
7. J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the 9^{th} ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677. ACM, 1998.
8. D. Lungley and U. Kruschwitz. Automatically maintained domain knowledge: Initial findings. In *Proceedings of ECIR*, pages 739–743, 2009.
9. C. Manning, R. Prabhakar, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.
10. K. Markey. Twenty-five years of end-user searching, Part 1: Research findings. *Journal of the American Society for Information Science and Technology (JASIST)*, 58(8):1071–1081, June 2007.
11. M. Mat-Hassan and M. Levene. Associating Search and Navigation Behavior Through Log Analysis. *JASIST*, 56(9):913–934, 2005.
12. F. Silvestri. *Mining Query Logs: Turning Search Usage Data into Knowledge*. Foundations and Trends in Information Retrieval. Now Publisher, 2009. Forthcoming.
13. J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.
14. M. White. *Making Search Work: Implementing Web, Intranet and Enterprise Search*. Facet Publishing, 2007.
15. R. W. White, M. Bilenko, and S. Cucerzan. Studying the Use of Popular Destinations to Enhance Web Search Interaction. In *Proceedings of SIGIR'07*, pages 159–166, Amsterdam, 2007.
16. R. W. White and I. Ruthven. A Study of Interface Support Mechanisms for Interactive Information Retrieval. *JASIST*, 57(7):933–948, 2006.