On the Feasibility of Supporting Encyclopedia Navigation with Proactive Search

Miikka Miettinen

Helsinki 9th December 2004

UNIVERSITY OF HELSINKI
Department of Computer Science
Research Seminar on Intelligent Information Retrieval

1 Introduction

According to Tennenhouse [Ten00], a system is proactive if it takes actions on behalf of the user without being under explicit control. The system should be able adapt to the individual preferences of the user, as well as the diversity of situations in which it participates. Technology should adapt to people rather than the other way round.

The idea of computers taking independent actions and making smart remarks is intuitively appealing, but very difficult to realize in practice. In this paper we explore the feasibility of retrieving relevant pages proactively from an online encyclopedia. The user is assumed to be looking for a specific piece of information, which is unknown to the system, but *partially observable* in the navigation and scrolling patterns emerging during the course of the task. In particular, the system should be able to identify fragments of text that draw the user's attention. Those fragments would then serve as a basis for generating queries to an underlying search engine. We would be mainly concerned with *similarity-based search*, trying to sample the set of potentially relevant documents the user might not be aware of. The results could be presented as an automatically updated list in a peripheral area of the user's display.

The next section introduces some relevant concepts and ideas from previous research. After that we move on to analyzing the particular characteristics of encyclopedia navigation as a potential application area for proactive search. Section 4 discusses the main components that a working system might have. However, it is impossible to discover the actual nature of the problem without empirical experiments. Section 5 describes a small-scale feasibility study, the primary purpose of which was to obtain some realistic data for testing the underlying ideas and assumptions. Section 6 concludes with the next steps that might be taken in this work.

2 Background and related work

Navigating a large collection of interlinked hypertext documents is a cognitively demanding task [Con87]. The users must simultaneously evaluate the information visible at a particular moment, recall the content and location of previously seen pages, and try to navigate systematically to an appropriate direction. Navigation itself involves locating links on the visible page, assessing their potential relevance,

and choosing one or more promising paths to explore to a variable depth. As a result, people quite often get confused and suffer from cognitive overload [Boe]. In subjective reports they typically associate poor performance with "feeling lost", and not knowing where to go and how to get there [McL89, EH89].

In certain types of document collections, including online encyclopedias, extensive linking is a key issue contributing to the utilization of the material. Moving from one potentially relevant page to another by locating the relevant links is in principle a very natural and effective way of satisfying certain kinds of information needs. The authors of the material cannot possibly predict the needs of a diverse audience, however, and have no choice but to provide the maximum number of links. As a result, the link structure of a large document collection is often very complex, and the risk of negative side effects, such as disorientation and cognitive overload, is real.

One proposed solution to this problem is *adaptive navigation support* [Bru01]. The goal is to create a simplified link structure, while retaining the kind of navigational freedom characteristic of hypertext content. The most common ways of approaching the problem are *guidance* and *orientation support*, both of which can be either *global* or *local* in nature [Bru96]. Global guidance means that the system suggests navigation paths through parts of the system, providing e.g. guided tours. In contrast, local guidance involves suggesting only the next step, often with a special link or button that takes the user to the page the system considers most appropriate to be visited next. The idea of orientation support is to present an overview of the link structure of a set of pages. Orientation support is considered global when the overview covers the entire system, whereas a local view shows only the pages that are in the "neighbourhood" of the currently displayed one.

Adaptive navigation support can be based on manipulation of the existing links as well as generation of entirely new links to create personalized navigation paths. The main forms in which the functionality of existing systems is provided to the user are direct guidance, link sorting, link hiding, and link annotation [Bru01]. The techniques are not mutually exclusive, and some systems employ several of them together.

Implementations of *direct guidance* recommend the one node, which is considered the best for the user to visit next. The link can be outlined visually or presented as an additional dynamic link. This kind of adaptive navigation support is very easy and convenient to use, but provides no help if the user does not want to follow the system's suggestion.

In *adaptive link sorting* the order of the links is chosen according to their estimated relevance to the user. Sorting is very common in various forms of information retrieval, but has limited utility in other applications. Most links are embedded in indexes, text or maps, and cannot be reordered without losing the context. Instability in menus has also been observed to cause usability problems, especially for people who are inexperienced with computers [DRD94, Kap93].

The idea of *adaptive link hiding* is to reduce the complexity of the hyperspace by showing only the most relevant links. Hiding is more transparent to the user than sorting, but makes the hidden items completely unavailable.

Adaptive link annotation involves attaching some kind of visual queues to the links to inform the user about their estimated relevance. Various kinds of symbols have been employed, including textual labels, icons, colors, and font sizes. Ordinary Web browsers implement a simple kind of adaptive annotation by showing visited links in different color. Link annotation provides functionality that is similar to sorting and hiding, while avoiding many of their problems.

In case the possible navigation paths are not restricted to links added by the authors of the material, adaptive navigation support can also appear in the form of *proactive search*. For example, the user could be supplied with links to other documents that discuss the same or related topics as the one that is currently visible [Bru01]. Rhodes and Maes [RM00] use the notion of Just-in-time Information Retrieval Agents (JITR agents) to characterize this type of a situation.[1] JITR agents "proactively retrieve and present information based on a person's local context in an easily accessible yet nonintrusive manner". The essence of the concept is the proactive but nonintrusive behavior of the software. Retrieving information proactively allows the agent to do it more often and with broader scope compared to an equivalent user-driven system. The majority of the suggestions may be irrelevant to the user, but as long as they can be easily ignored, the cost of the mistakes is supposedly outweighed by the occasional success. Based on small-scale experiments with three different applications, Rhodes and Maes claim that users equipped with their JITR agents end up accessing and utilizing more information than they would if they had to use the retrieval facilities themselves.

Many existing implementations of adaptive navigation support rely on handcrafted *adaptation rules* [Bru01, DC98]. Such rules can be expressed in several ways. In

[1] Although the implementations discussed by Rhodes and Maes are not directly related to adaptive navigation support, the *concept* is useful and relevant for the present discussion.

some systems they are attached directly to individual pages. For example, the author could provide a rule stating that "if the user is a novice, recommend this additional explanation". Typically, the user's level of expertise would either be asked from him directly or estimated on the basis of his previous interactions with the system. The number of rules needed can often be decreased by specifying them with respect to classes of pages instead of individual pages. Some kind of a sign could be attached to all "additional explanations for novices", and a single adaptation rule applied to all of them. The third alternative is to rely on a structural description of the content. In some educational systems, for example, handcrafted knowledge of prerequisite relationships is combined with navigation data [BS98]. If the user has not studied all of the pages that provide the background for a particular link, the system is able to indicate the appropriate order of studying the pages. The advantage of this approach is that individual variations in navigation history are relatively easy to take into account. In contrast, explicit rules often need to be simplified by grouping the users into a small number of stereotypes.

Another approach that can be used for providing adaptive navigation support is content-based filtering. The idea is to maintain an *interest profile* for the user based on his or her previous interactions with the system, and apply the profile to making predictions of the relevance of other available pieces of content [MGT+87]. The underlying assumption is that the information needs of the user can be characterized adequately by a single profile that is reasonably stable across different sessions and longer periods of time. The data used for constructing the profile may consist of explicit ratings or implicit feedback accumulating as a side product of the normal interactions with the system (e.g. click-stream data) [AY00]. In order to generalize the profile to items that the user has not seen, the system needs to employ either a suitable categorization scheme or some kind of a model of the contents. In the first case it is typically assumed that a user who has indicated interest in e.g. basketball might want to see additional items associated with that category. Models of text documents are normally based on either vector space representations with TF-IDF weighting [AY00] or statistical models such as mPCA [Bun02] or latent Dirichlet allocation [BNJ03].

The third basic approach is to take advantage of statistical regularities within groups of users. The first major class of applications based on this idea was *recommender systems* (which are nowadays better known as *collaborative filtering* systems) [RV97]. In their most basic form, they ask the user to rate a number of familiar items, after which they provide a list of additional items that other users with similar preferences

have liked. The items can be e.g. music albums [SM95], movies [BHK98], newsgroup messages [KMM+97], or restaurants [BHY97]. Some attempts have been made to use navigation data instead of ratings to eliminate the extra work involved in providing explicit inputs to the adaptation mechanism (see e.g. [WLL+03, MCS00]).

Collaborative filtering systems rely on statistical modeling, and need fairly large amounts of data to be able to make reliable predictions. This often causes certain practical problems, including the *cold start problem* and the *sparsity problem* [SPUP02]. The cold start problem appears both at the beginning of a system's lifetime and whenever newly added items have not accumulated enough ratings or implicit relevance data. On the other hand, the sparsity problem is an issue when the total number of items is orders of magnitude larger than the number of ratings made by a single individual. In this case insufficient overlap in the ratings of the individual users prevents the system from making appropriate generalizations. One reasonable solution is to design the system so that it can function as a normal hypertext or information retrieval system in the beginning, and gradually increase its level of adaptivity as more data becomes available [BB02]. Collaborative filtering has also been combined with content-based filtering in order to generalize the relevance feedback over both users and items (see e.g. [CGM+99, BH04])

Finally, adaptive navigation support could be built on top of a search engine. The key underlying feature required for this is *similarity-based search*. Both vector space and statistical models support a notion of distance between bags of words (which is the way the documents are represented in the search engine index). The idea would be to generate a suitable bag of words automatically on the basis of the users' past and present navigation patterns, and send it as a query to the underlying search engine. Unlike the other approaches, this would enable dynamic switches of focus, as the query could be based on both the momentary situation and the longer-term history of the user.

3 Adaptive navigation support in an online encyclopedia

Online encyclopedias have some particular characteristics, which limit the applicability of the techniques discussed in the previous section. In terms of content, they are very broad and shallow. They try to cover superficially almost every topic that

the users might possibly be interested in. As a result, the number of different information needs that are being addressed by a comprehensive online encyclopedia is enormous.

From the user's point of view, an encyclopedia is primarily a resource for finding an answer to a specific question. In other words, the navigation of the user is typically oriented towards a particular goal.[2] However, it is reasonable to assume that the goal can never be observed directly by the system.

The broad scope and superficial nature of the material along with the way people utilize it makes the notion of a user profile irrelevant. The users do not have a fixed set of interests or preferences that would characterize their navigation patterns adequately. Rather, they turn to the encyclopedia driven by wide variety of highly specific needs. This makes rule-based techniques and collaborative filtering inappropriate for providing adaptive navigation support, since they both rely on the assumption that the users' needs are focussed on an identifiable subset of the material. In contrast, content-based techniques do not necessitate this assumption, as long as the underlying models can be utilized with sufficient flexibility.

Typing in the search terms that locate the right document directly is always faster and more convenient than navigating along links. For this reason, adaptive navigation support is relevant only when the user is unable to come up with a sufficiently distinctive set of search terms. This typically happens when identifying them requires specific knowledge of the domain, or when the relevant information is not associated with any distinctive vocabulary. In such situations a natural strategy is to resort to hierarchical navigation. The user might start from a general page related to the topic of interest, trying to find promising links or better search terms. To the extent that this strategy is successful, the user will move gradually towards increasingly specific pages, eventually finding the right piece of information.

Navigating complex link structures is a demanding task, however, as was pointed out in the previous section. In addition, it may not be clear to the user what kind of pages exist and which paths need to be followed to reach them. Adaptive navigation support could help by providing shortcuts and making it easier for the user to focus the navigation to the right direction.

The idea that we will explore in more detail in the next section is to provide this

[2] Of course, casual browsing is another common way of using an online encyclopedia. The effectiveness of adaptive navigation support would be very difficult to evaluate in that setting, however, and it is therefore not considered in this paper.

kind of functionality in the form of proactive search. As the user engages in hierarchical navigation towards the relevant information, the system would observe his or her click stream and scrolling patterns and retrieve potentially relevant links from an underlying search engine. According to Rhodes and Maes [RM00], the precision of proactive search does not need to be very high if the results are presented in an unobtrusive manner. In a sense, the system would be *sampling* a set of potentially relevant pages in an attempt to provide the user with a selection of possible navigation paths.

4 Elements of a proactive search engine

As described in the previous section, the task of the proactive search facility would be to support encyclopedia navigation by retrieving potentially relevant links on the basis of the user's observed navigation and scrolling patterns.[3] Figure 1 shows the overall architecture of the hypothetical system. Between the user's web browser and the other parts of the system is a proxy server, which takes care of analyzing the data and sending queries to a topic-based search engine. The particular search engine that we will use in our future experiments is described in [BLP+04]. The proxy server also intercepts requests for encyclopedia content in order to accumulate the navigation data and augment the pages with the additional machinery needed for tracking the scrolling patterns of the user.[4]

The chain of events resulting to an updated list of recommended links is triggered from the browser (see Figure 2). The first step towards generating the search engine query is to retrieve relevant parts of the user's scrolling history from a database. It is probably sufficient and appropriate to look at the data associated with the current page and a few preceding pages in order to account for the dynamic nature of the information seeking process.

There are two slightly different perspectives to the utilization of the scrolling data. The more straightforward one is to treat the amount of time that a particular fragment of text has been visible on the user's screen as a weight for the words contained in the fragment. The generation of the query would (in the simplest implementation)

[3] We provide just a very high-level overview of the key components that might be needed for handling the task. Lots of essential details are left unspecified, because a sufficient understanding of the problem has not been achieved yet. The main purposes of this section are to illustrate the ideas presented in other parts of the paper and to provide a basis for discussion.

[4] Contrary to a common belief, this can be done easily with some standard JavaScript.

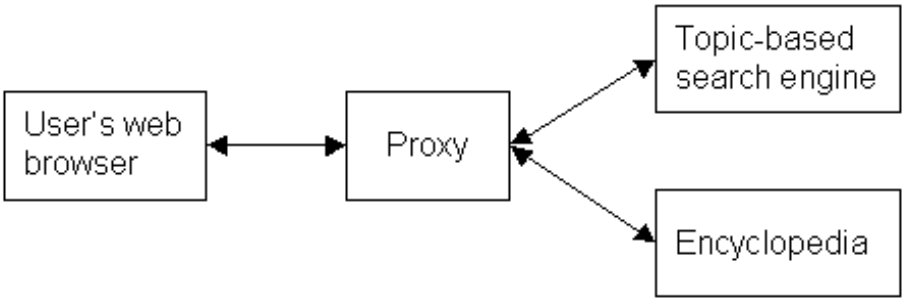

Figure 1: High level architecture of the system.

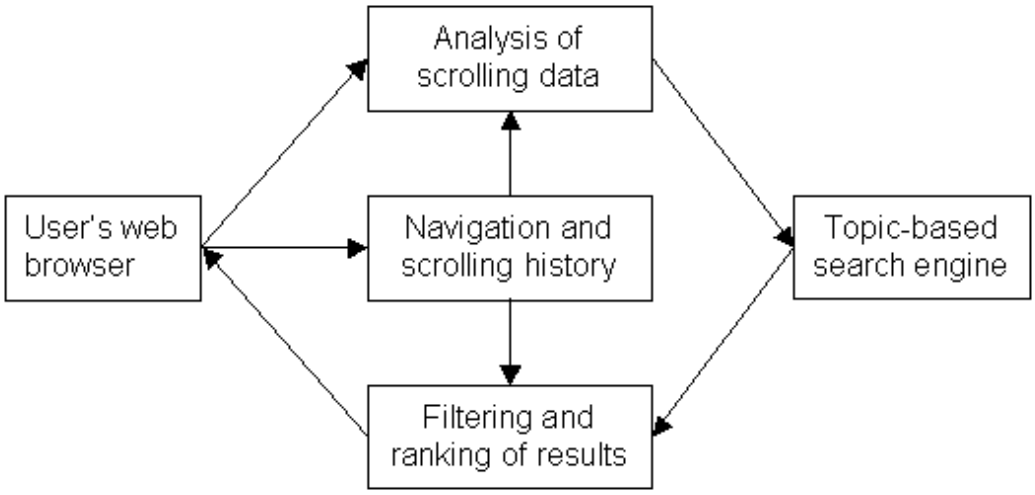

Figure 2: Proactive retrieval of links.

involve creating a bag of words with a composition reflecting the weights. The other approach is to remove from the query the pages and page fragments that the user apparently has not considered relevant. This could be done by making a distinction between "reading" and "browsing" on the basis of the observed scrolling patterns and the layout of the page. A Hidden Markov Model [Rab89] learned from previously gathered data might be capable of making this kind of a distinction. In either case, a query is sent to the search engine to retrieve a list of documents that are similar to the material the user has been viewing.

Before the results are sent back to the browser, it is appropriate to remove at least the pages that the user has already visited. It might also be worthwhile to consider prioritizing the list somehow. If the resolution of the statistical model used by the search engine is sufficient, it might be possible to estimate the degree of specificity of a particular document, and try to account for the user's current position in the (supposedly) hierarchical navigation process.

5 Feasibility study

Due to the nature of the problem, it would be very difficult to build a working system without some real data of the users' navigation and scrolling patterns. Rather, the optimal development process would involve building a series of prototypes, and evaluating them with subjects engaging in realistic information seeking tasks. As a step in this direction, we conducted a small-scale feasibility study in a semi-controlled setting. Besides providing us with some test data for developing the first versions of the proactive search, the experiment was useful for gaining a better intuitive understanding of the issues involved.

An attempt was made to come up with a set of reasonable information seeking tasks that would require some hierarchical navigation from general pages to more specific ones. However, it was emphasized in the instructions that the subjects were allowed to use any suitable methods for finding the answer, including in particular the available facilities for keyword search. Our intention is to explore an additional piece technology that should fit in the existing framework. Therefore, it would be unrealistic to ignore the existence of keyword search and assume navigation to be the only method of finding information.

As the information source, we used a locally installed version of Wikipedia.[5] The subjects were given the following 10 questions along with instructions to find the right answer from Wikipedia:

1. What are the distinctive characteristics of the architecture of the first half of the 20th century?
2. Do insects breathe?
3. What ethical problems would be involved in transforming Mars (or some other planet) into an environment suitable for a permanent human population?
4. Which type of piece in Western classical music relies particularly heavily on turn-taking and "dialogue" of melodies?
5. What techniques are available for monitoring the health of a fetus?
6. What kind of social order was prevalent in most countries of Western Europe during the Middle Ages?
7. How does the device used for measuring air pressure work?
8. What kind of analytical methods do companies use when making investment decisions?
9. Which international court deals with human rights?
10. Which force keeps the atomic nucleus together?

The questions were presented one at a time in a random order. Figure 3 shows the appearance of the user interface. Wikipedia was shown inside a fixed size inline frame on the left to eliminate variations in scrolling data that could have been caused by differing screen resolutions. The ongoing task was shown on the right in both Finnish and English. The page containing the answer was supplemented with a special button, which allowed the user to register the answer and move to the next task. It was also possible to skip a task by clicking the "Next task" link.

The study was conducted during 6.12.-7.12.2004 with 7 subjects, who in total completed 45 tasks successfully. Due to severe limitations in space and time, I will discuss the results only in the oral part of my presentation.

[5] See http://en.wikipedia.org.

Figure 3: The user interface of the experiment.

6 Conclusions and future work

Whether or not it is feasible to support encyclopedia navigation with proactive search is still an open issue that we will continue exploring. Our next step is to analyze the available data and develop the first versions of the proactive search functionality. If the results are encouraging, we will conduct a larger scale controlled experiment, in which we will try to quantify the benefit of this type of navigation support in a task similar to the feasibility study.

References

AY00 Aggarwal, C. C. and Yu, P. S., Data mining techniques for personalization. *IEEE Data Engineering Bulletin*, 23,1(2000), pages 4–9.

BB02 Baudisch, P. and Brueckner, L., TV Scout: lowering the entry barrier to personalized TV program recommendation. *Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer-Verlag, 2002, pages 58–67.

BH04 Basilico, J. and Hofmann, T., Unifying collaborative and content-based filtering. *Proceedings of the twenty-first international conference on machine learning*. ACM Press, 2004.

BHK98 Breese, J. S., Heckerman, D. and Kadie, C., Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 1998, pages 43–52.

BHY97 Burke, R. D., Hammond, K. J. and Young, B. C., The FindMe approach to assisted browsing. *IEEE Expert*, 12,4(1997), pages 32–40.

BLP+04 Buntine, W., Löfström, J., Perkiö, J., Perttu, S., Poroshin, V., Silander, T., Tirri, H., Tuominen, A. and Tuulos, V., A scalable topic-based open source search engine. *Proceedings of the IEEE/WIC/ACM Conference on Web Intelligence (WI 2004)*, 2004, pages 228–234.

BNJ03 Blei, D. M., Ng, A. Y. and Jordan, M. I., Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3, pages 993–1022.

Boe Boechler, P. M., How spatial is hyperspace? interacting with hypertext documents: Cognitive processes and concepts. *Cyberpsychology & Behavior*, 4, pages 23–46.

Bru96 Brusilovsky, P., Methods and techniques of adaptive hypermedia. *User Modeling and User Adapted Interaction*, 6,2–3(1996), pages 87–129.

Bru01 Brusilovsky, P., Adaptive hypermedia. *User Modeling and User Adapted Interaction*, 11,1–2(2001), pages 87–110.

BS98 Brusilovsky, P. and Schwarz, E., Web-based education for all: A tool for developing adaptive courseware. *Computer Networks and ISDN Systems*, 30,1–7(1998), pages 291–300.

Bun02 Buntine, W., Variational extensions to EM and multinomial PCA. *Proceedings of the 13th European Conference on Machine Learning*, 2002, pages 23–34.

CGM+99 Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. and Sartin, M., Combining content-based and collaborative filters in an online newspaper. *Proceedings of the ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*. ACM Press, 1999.

Con87 Conklin, J., Hypertext: An introduction and survey. *Computer*, 20,9(1987), pages 17–41.

DC98 DeBra, P. and Calvi, L., AHA: a generic adaptive hypermedia system. *Proceedings of the 2nd Wordshop on Adaptive Hypertext and Hypermedia (Hypertext'98)*, 1998, pages 5–11.

DRD94 Debevc, M., Rajko, S. and Donlagic, D., Adaptive bar implementation and ergonomics. *Informatica: Journal of Computing and Informatics*, 18, pages 357–366.

EH89 Edwards, D. W. and Hardman, L., Lost in hyperspace: Cognitive mapping navigation in a hypertext environment. In *Hypertext: Theory into practice*, McLeese, R. A., editor, Ablex, Westport, CT, 1989, pages 90–105.

Kap93 Kaptelinin, V., Item recognition in menu selection: The effect of practice. *INTERCHI'93 Adjunct Proceedings*, New York, NY, USA, 1993, ACM Press, pages 183–184.

KMM+97 Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L. and Riedl, J., Grouplens: Collaborative filtering for usenet news. *Communications of the ACM*, 40,3(1997), pages 77–87.

McL89 McLeese, R. A., Navigation and browsing in hypertext. In *Hypertext: Theory into practice*, McLeese, R. A., editor, Ablex, Westport, CT, 1989, pages 6–44.

MCS00 Mobasher, B., Cooley, R. and Srivastava, J., Automatic personalization based on web usage mining. *Communications of the ACM*, 43,8(2000), pages 142–151.

MGT+87 Malone, T. W., Grant, K. R., Turbak, F. A., Brobst, S. A. and Cohen, M. D., Intelligent information sharing systems. *Communications of the ACM*, 30,5(1987), pages 390–402.

Rab89 Rabiner, L. R., A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77,2(1989), pages 257–286.

RM00 Rhodes, B. J. and Maes, P., Just-in-time information retrieval agents. *IBM Systems Journal special issue on the MIT Media Laboratory*, 39,3–4(2000), pages 685–704.

RV97 Resnick, P. and Varian, H. R., Recommender systems. *Communications of the ACM*, 40,3(1997), pages 56–58.

SM95 Shardanand, U. and Maes, P., Social information filtering: Algorithms for automating "word of mouth". *Proceedings of the 1995 Conference on Human Factors and Computing Systems (CHI 1995)*. ACM Press, 1995, pages 210–217.

SPUP02 Schein, A. I., Popescul, A., Ungar, L. H. and Pennock, D. M., Methods and metrics for cold-start recommendations. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, 2002, pages 253–260.

Ten00 Tennenhouse, D., Proactive computing. *Communications of the ACM*, 43,5(2000), pages 43–50.

WLL+03 Wettig, H., Lahtinen, J., Lepola, T., Myllymäki, P. and Tirri, H., Bayesian analysis of online newspaper log data. *Proceedings of the 2003 Symposium on Applications and the Internet Workshops (SAINT 2003 Workshops)*. IEEE Computer Society Press, 2003, pages 282–287.