18

Adaptive News Access

Daniel Billsus[1] and Michael J. Pazzani[2]

[1] FX Palo Alto Laboratory, 3400 Hillview Ave., Bldg. 4,
Palo Alto, CA 94304 , USA
billsus@fxpal.com
[2] Rutgers University, ASBIII, 3 Rutgers Plaza
New Brunswick, NJ 08901
pazzani@rutgers.edu
pazzani@rutgers.edu

Abstract. This chapter describes how the adaptive web technologies discussed in this book have been applied to news access. First, we provide an overview of different types of adaptivity in the context of news access and identify corresponding algorithms. For each adaptivity type, we briefly discuss representative systems that use the described techniques. Next, we discuss an in-depth case study of a personalized news system. As part of this study, we outline a user modeling approach specifically designed for news personalization, and present results from an evaluation that attempts to quantify the effect of adaptive news access from a user perspective. We conclude by discussing recent trends and novel systems in the adaptive news space.

18.1 Introduction

The World Wide Web has had a profound impact on our everyday lives: we routinely rely on it as a ubiquitous source for timely information. In particular, the web's real-time and on-demand characteristics make it an ideal medium for news access anywhere and anytime. As a result, virtually every news organization now has a presence on the World Wide Web.

In addition to transforming how traditional news organizations distribute information, the web has enabled new forms of information dissemination. The web provides an audience and a platform for individuals to express themselves or engage in discussions. Weblogs, or *blogs*, on thousands of topics contributed by thousands of individuals have created a rich information landscape of gigantic proportions.

While the availability of continuously updated news content provides great value, it represents yet another facet of the, now omnipresent, information overload problem. How can individuals find the most interesting or relevant news content? How can they discover trusted news organizations or find relevant blog posts among thousands of choices?

Adaptive web technology provides the basic building blocks to address these challenges. We now know how to build tools that help people discover relevant content,

P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.): The Adaptive Web, LNCS 4321, pp. 550 – 570, 2007.

route the right information to the right people at the right time, or help aggregate content from thousands of sources.

In this section, we show how some of the adaptive web technologies discussed in previous chapters have been applied to adaptive news access. First, we provide an overview of different types of adaptivity in the context of news access and identify corresponding algorithms. For each adaptivity type, we briefly discuss representative systems that use the described techniques. Next, we discuss an in-depth case study of a personalized news system. As part of this study, we outline a user modeling approach specifically designed for news personalization, and present results from an evaluation that attempts to quantify the effect of adaptive news access from a user perspective. We conclude by discussing recent trends and systems in the adaptive news space.

18.2 Types of Adaptive News Access

The main goal of adaptive news techniques is to facilitate access to relevant news content. This goal can be achieved in several different ways. In this section, we take a closer look at the following types of adaptivity:

- *News Content Personalization*. Systems that personalize content help users find personally relevant news stories based on a model of the user's interests. These systems can recommend or automatically rank stories, so that the most relevant content is easier to find.
- *Adaptive News Navigation*. Adaptive navigation assists the user in navigating to the most frequently read sections of a news site.
- *Contextual News Access*. Contextual news access techniques provide users with news content on the basis of currently viewed information.
- *News Aggregation*. Automated aggregation and classification of news content helps users identify ongoing or emerging news topics, and assists in accessing coverage of a specific topic by multiple providers. In contrast to the adaptivity types listed above, news aggregation does not necessarily enable personalization, but automated aggregation commonly exhibits adaptive behavior: dynamically generated aggregator pages, e.g. Google News, automatically adapt to the current news landscape and provide an indication of emerging topics or trends.

While this chapter focuses on adaptive news access, it is important to point out that the analysis of news content for purposes other than adaptivity has been the focus of much research in the information retrieval and machine learning communities. For example, Topic Detection and Tracking (DTD) is a DARPA-sponsored initiative to investigate techniques for finding and following new events in streams of broadcast news stories ([35] provides an overview of the DTD sub-tasks and corpora). Clearly, research results from these areas are often directly applicable to adaptive news access techniques.

18.2.1 News Content Personalization

This section focuses on adaptive techniques that model the user's interests based on explicit or implicit feedback, and use the resulting user models to personalize news content. Collecting user feedback explicitly or implicitly has been the focus of much research. See Chapter 2 of this book [13] for an in-depth discussion of various feedback methods. In addition, the value and accuracy of implicit user feedback have been studied in the context of news personalization ([2], [21], [17], [18]).

The adaptive techniques described in this chapter are in contrast to static content customization via user-provided interest profiles. For example, many major news sites, e.g. Yahoo! News or Google News, allow users to customize the news categories to be included on the front page, or allow users to indicate interesting topics via web questionnaires. To distinguish between these two approaches, we refer to user-defined news profiles as *customization* and adaptive techniques as *personalization*.

Previous sections of this book discussed techniques that model users' individual interests to personalize content (see Chapters 2, 3, 9 and 10 of this book [13], [26], [30], [27]). Many of the described modeling and recommendation techniques apply directly to news personalization. However, news access has several characteristics that make some approaches better suited to the problem than others. In this section, we describe how news access differs from other personalization tasks and suggest a set of techniques that are particularly appropriate for this domain. The following characteristics are important factors for the design of adaptive approaches to news personalization.

Dynamic Content. News content is more dynamic than many other content types, such as movies, music or books. News stories are released and updated continuously, and many stories only remain online for a short period of time until new details emerge. This makes content-based methods better suited to news personalization than collaborative methods. As discussed in Chapter 9 of this book [30], collaborative filtering is based on using the ratings of like-minded users as predictors for relevant content. However, collaborative filtering often suffers from the "sparse matrix" problem, and this limitation is particularly noticeable for news access. For example, ratings from users who have accessed a story can often not be used as predictors for the current user because of very limited rating overlap. Related to this issue, collaborative filtering approaches suffer from a "latency" problem, i.e. depending on a site's popularity and traffic, it may take some time for news stories to receive enough user feedback to lead to accurate recommendations. In contrast, content-based methods predict the user's interest using text alone, and do not depend on the availability of ratings. In addition, the benefits of collaborative filtering, i.e. being able to take advantage of qualitative human judgments, are often not critical for services that serve news from a single provider. Once the user has selected a content provider he or she trusts and agrees with ideologically, the selection of relevant stories is primarily an issue of content and not quality or style. However, collaborative methods can certainly be very useful for services that attract a large number of users or aggregate stories from multiple providers. For example, the *GroupLens* system is well known for its application of collaborative filtering techniques to Usenet news [20].

Changing Interests. Since new news topics emerge continuously, users' interests tend to change frequently. From an algorithmic perspective, this calls for methods that can quickly adjust to changing target concepts. For example, a system that uses a machine learning algorithm to learn a model of the user's interests (as discussed in Chapter 10 of this book [27]), should be based on learning algorithms that can quickly adjust to changing interests, so that the user does not have to provide a large number of training examples until the system discovers the interest change. More generally, the notion of changing target concepts that must be tracked algorithmically is known as concept drift. The machine learning literature discusses many approaches specifically designed to address this problem. These techniques are often based on algorithms that can either explicitly detect changing concepts or limit the effects of concept drift via windowing techniques that only consider a temporally constrained subset of the available data [36], [19]. In addition, learning algorithms that require only a few training examples to approximate useful models can quickly pick up new user interests. For example, instance-based algorithms such as the *k*-nearest-neighbor algorithm fall into this category ([10], Chapter 10 of this book [27]).

Multiple Interests. Users are usually interested in a broad range of different news topics. This means that a user modeling approach for news must be capable of representing multiple topics of interest. If the user model is inferred by a machine learning algorithm, *k*-nearest-neighbor methods are a good choice to address this issue. Suppose the user model is based on the *k*-nearest-neighbor-based representation described in Chapter 10 of this book [27]: in this case, the user model contains labeled news stories, e.g. interesting vs not interesting, where each story is represented as a vector in the Vector Space Model [33]. A previously unseen story can now be classified using the labels of the *k* most similar stories in the user model. Since the classification only depends on these *k* nearest neighbors, stored news stories that are "further away" from the story to be classified and likely reflect the user's interests in different topics, do not influence the classification. As a result, *k* nearest neighbor methods lend themselves to modeling multiple disjoint areas of user interests.

Novelty. A news story is usually considered most interesting if it conveys information the user does not yet know. This has interesting algorithmic implications and further differentiates news access from other domains where finding "more of the same" can be a good thing. Algorithms that keep track of information the user has previously accessed can avoid this problem by selecting content that is similar, but not identical to, previously accessed information. For example, Newsjunkie [12] is a system that personalizes news for users by identifying the novelty of stories in the context of stories they have already reviewed. The system uses novelty-analysis algorithms that analyze inter- and intra-document dynamics by considering how information evolves over time from article to article, as well as within individual articles.

Avoiding Tunnel Vision. Clearly, personalization should not prevent the user from finding important novel information or breaking news stories. Adaptive news systems can overcome this issue by optionally integrating editorial input into the recommendation algorithm, or by explicitly boosting the diversity of stories presented to the user (see Section 18.3 for an example and more details).

Editorial Input. Adaptive news systems attempt to identify a set of stories deemed most "relevant" within a large set of potential candidate stories. Traditionally, this has been the job of a news editor, i.e. a person who decides which stories to include in a paper (or site), and how to order and categorize them. The advent of adaptive news technology does not imply that human news editors are not needed anymore. Ranking stories by their perceived importance is hard to automate and, ideally, the input from human editors should inform the automated selection of content via learned user models. In addition to obvious benefits from a user perspective, retaining editorial input is an important feature for news organizations that are interested in deploying personalization technology: loss of control over the content that users will get to see does not appeal to news organizations. Retaining editorial input can be achieved via simple methods, such as factoring the position of a news story (as a measure of its relative importance) into recommendation algorithms, or by specifying selection rules that ensure the user will always get to see the top *n* stories, regardless of the user's interest model [5]. This is an issue that primarily concerns news personalization for individual news organizations. For systems that aggregate news content from multiple providers this is less of an issue, and, in fact, one goal of such aggregation systems can be to overcome editorial bias.

Brittleness. A single action, such as selecting something accidentally or skipping over an article on a topic (perhaps because one heard about the details on the radio or the wireless connection is dropped) should not have a drastic or unrecoverable effect on the model of the user's interests.

Availability of Meta-Tags. News personalization algorithms can usually not rely on the availability of meta-tags. A process that requires the content provider to do extra work by hand, such as adding meta-tags or category labels, is not feasible with thousands of new items being added daily.

In Section 18.3, we outline one particular algorithm that addresses most of the characteristics identified above in more detail, and we highlight results from an experimental evaluation. In addition to the algorithm described in Section 18.3 and the examples listed above, there are many alternative ways to address the issues identified in this section, and numerous studies that discuss various facets of content-based news personalization have been published. For example, early work by Bharat et al. [2] focuses on a dynamic user interface for personalized web-based news access: according to the authors, the *Krakatoa Chronicle* was the first newspaper on the World Wide Web to provide a layout similar to that of real-world newspapers. The system is highly interactive, supports article layout customization, and provides a content-based personalization approach that allows users to control the extent to which public and personal interests affect the selection of articles. Closely related to this work, Sakagami and Kamba [29] describe the *Anatagonomy* system, a research prototype of a personalized online newspaper. This work mainly focuses on the utility of implicit vs explicit user feedback. For example, *Anatagonomy* tracks user interactions such as selecting and enlarging articles or scrolling through articles, and interprets these interactions as implicit feedback. Similarly, Balabanovic [1] describes *Slider*, a user interface specifically designed to capture users' preferences implicitly, in the context of content-based news personalization. The *Slider* interface is based on a set of on-screen panels that users can associate with topics they find interesting. At any time, users can create new panels or delete old panels in order to reflect their changing interests. When new stories arrive, users can optionally "slide" these stories onto a panel, and thereby implicitly define the topic associated with the panel. News stories are represented in the Vector Space model [33], and for each panel, the system constructs one prototype vector by averaging over all documents contained in the panel. In subsequent sessions, the system attempts to locate news stories that are similar to these prototype vectors.

In general, purely statistical Information Retrieval techniques (such as the Vector-Space-model-based approaches described in this chapter) are commonly used as the underlying foundation for content-based news personalization approaches. However, there has also been work that examines the utility of linguistic information in the context of news personalization. For example, Magnini and Strappavara [23] explore the utility of word-sense information for user profile acquisition, and report a tangible increase in recommendation accuracy, compared to a purely statistical approach. Additional content-based news personalization studies of interest include [14], [21], [16], [32] and [22].

Finally, it is important to point out that news content personalization systems have not only created academic interest, but are starting to become publicly available as part of commercial news services. An early example of a publicly available personalized news service that automatically learns from users' access patterns is Findory *(http://www.findory.com)*. For each individual user, Findory tracks accessed stories and uses this information to generate a personalized front page. For example, the page

Fig. 18.1. Findory's adaptively personalized front page. Recommended stories are annotated with a "sun" icon.

shown in Figure 18.1 is the result of selecting several stories about security issues and search technology (recommended stories are annotated with a “sun” icon). While the exact details of the personalization algorithm are proprietary, the Findory web site states that the “personalization algorithm combines statistical analysis of the article's text and behavior of other users with what we know about articles you have previously viewed”, which suggests that the system is based on a combination of content-based and collaborative methods.

18.2.2 Adaptive News Navigation

Similar to personalization based on content profiles, the goal of adaptive navigation is to simplify access to relevant content. However, instead of finding individual news stories that match the user’s interest profile, this technique focuses on analyzing the user’s access patterns to determine the position of menu items within a menu hierarchy. For example, a user who frequently accesses the technology section of a news site, but never reads sports stories, would probably prefer to see the technology category along with individual technology stories high up on the front page of a personalized news site. This personalization approach is particularly effective for mobile applications on PDAs and cell phones, because of the limited screen space of these devices. For example, Smyth and Cotter [31], describe an algorithm that personalizes the menu hierarchy of a mobile application based on menu access frequencies that are maintained for each individual user. The system estimates the probability that a user will select option *o* given that it is included in menu *m*, and uses these probabilities to construct menus that are most likely to contain options the user will select. An empirical evaluation of this approach applied to a mobile portal showed that, on average, the number of menu-select and scroll operations was reduced by over 50%, leading to a much improved user experience.

Since this approach does not use any content and is primarily geared towards adaptive menu reordering, it does not lend itself to recommendations for individual news stories. However, the simplicity of adaptive menu reordering is also its greatest strength: it does not require complex infrastructure that maintains large content-based profiles for individual users, which means that it is much easier to deploy and satisfy real-world scalability requirements than more complex techniques. In addition, the approach does not require a lengthy training period, leading to significant usability improvements even after only a single user session.

While Smyth and Cotter [31] demonstrate the approach in the context of a mobile portal, it can be applied to any application with long or complex menu hierarchies. For example, promoting a user’s favorite news categories close to the top of a personalized news site is a useful application of this approach, followed by several mobile news sites. For example, at the time of this writing, the San Diego Union Tribune maintained an adaptive news site for mobile devices that reorders its news categories based on access frequencies at *http://go.sosd.com*.

Clearly, adaptive navigation and content personalization are not mutually exclusive. A combined approach could provide access to the user’s favorite news categories based on access probabilities, and each accessed news category could contain a personalized set of stories based on a content-based user profile.

18.2.3 Contextual Recommendations

Contextual recommendations, sometimes referred to as just-in-time retrieval, are closely related to content-based personalization [28]. However, instead of using a model of the user's interests learned over time, the approach draws on currently displayed information, such as a web page or email message, as an expression of the user's current interests. For example, a contextual news recommender could recommend a news story about a certain company when the user visits a web page that contains information about the company. Likewise, a contextual news recommender could recommend a news story about an actor when the user receives an email message from a friend that mentions the actor's name.

Contextual recommenders typically operate as follows. First, the system extracts textual information currently displayed on the user's screen. This can be accomplished via plug-ins for commonly-used applications, such as web browsers, email clients or word processors. Alternatively, web proxies can be used to access the text currently displayed in the user's browser. The extracted text is then used to retrieve related content, such as related news stories. This step of the process is often based on statistical text processing: using statistical term-weighting techniques ([33], Chapter 5 of this book [25]) to identify informative terms, the text can be used to automatically construct a query which can subsequently be sent to a search engine to retrieve related content. In addition to statistically determined query terms, Natural Language Processing approaches can be used to assist in the query generation process. For example,

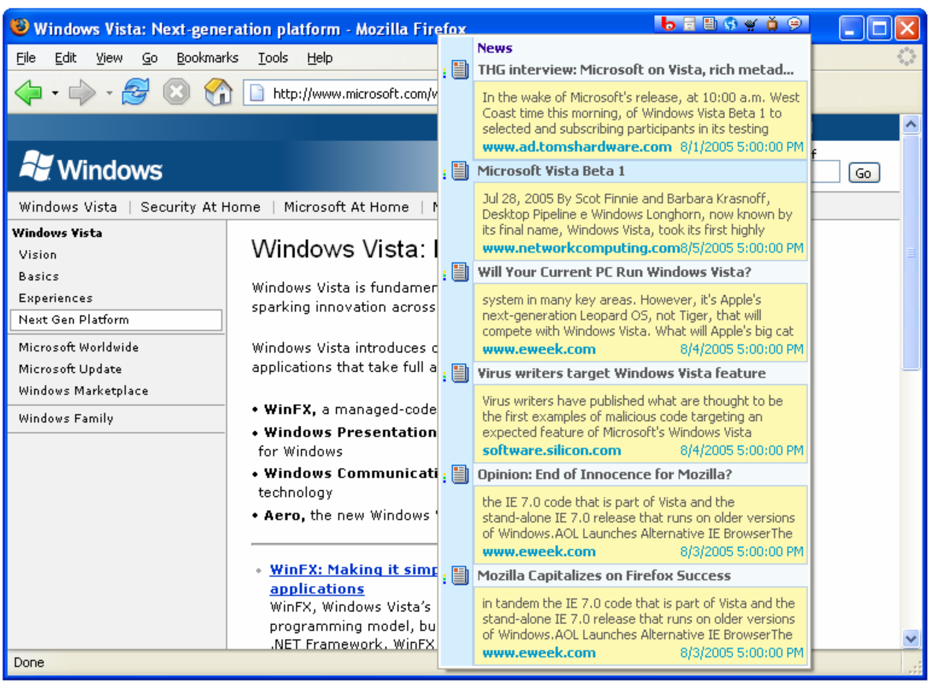

Fig. 18.2. Blinkx recommends news stories based on contextual information. In this example, Blinkx recommends news stories related to a web page the user is viewing.

a named-entity-tagger [7] can help identify names of companies, people or products for inclusion in the automatically generated query. *Watson* [8] and *FXPAL Bar* [6] are examples of this type of system. Both systems are implemented as toolbars that run within web browsers, email clients or other applications, and communicate the availability of related content via subtle icon cues, such as displaying a light bulb or changing the color of a button. Similarly, *Blinkx* is a publicly available contextual recommender (*http://www.blinkx.com*). Like *Watson* and *FXPAL Bar*, *Blinkx* is a toolbar that proactively recommends content in response to changing information on the user's screen. In addition to web pages, products and TV content, *Blinkx* recommends news stories. In the scenario shown in Figure 18.2, the user is viewing a web page about the next version of Microsoft Windows. The *Blinkx* bar in the upper right corner of the screen indicates the availability of related news stories via a small paper icon. When the user clicks on this icon, a list of closely related news stories appears (about Microsoft's new operating system, in this example).

18.2.4 News Aggregation

The adaptivity types discussed in the sections above focus on inferring users' interests based on the browsing history or currently viewed information. In contrast, news aggregators are services that automatically aggregate content from many different news sources and, as a result, adapt to the current news landscape as a whole. A technical trend that has significantly contributed to the emergence of news aggregators and other news-related services is the widespread use of RSS (Really Simple Syndication) feeds. A news or blog provider can publish RSS feeds, i.e. XML documents that provide links to currently available content, to simplify syndication. The XML excerpt below shows an example of a simple RSS feed.

```
<?xml version="1.0"?>
<rss version="2.0">
  <channel>
    <title>My Gaming News</title>
    <link>http://mygamingnews.com/</link>
    <description>Gaming News</description>
    <item>
      <title>600,000 Xbox 360 units sold in US</title>
      <link>http://mygamingnews.com/story1.html</link>
      <description>LOS ANGELES -- Microsoft Corp.
          has sold 600,000 of its new XBox 360
          videogame consoles …
      </description>
    </item>
    <item>
      <title>'Grand Theft Auto' slapped with lawsuit
      </title>
      <link>http://mygamingnews.com/story2.html</link>
    </item>
  </channel>
</rss>
```

For example, *Google News* (*http://news.google.com*) is a popular news aggregation service. The site currently collects stories from more than 4,500 news sources in English, automatically identifies common topics, ranks these topics by estimated "importance" (measured in terms of recency and volume), and then generates a new front page. This means that *Google News* adaptively generates a front page without explicit editorial input from a human editor, i.e. the aggregation acts as an unbiased news editor. While the exact details of the topic identification and ranking algorithm are proprietary, a news aggregation service can be implemented using statistical term weighting and text similarity techniques to automatically assess the similarity of any two stories ([33], Chapter 5 of this book [25]). In addition, text categorization approaches (see Chapter 10 of this book [27]) can be used to train classifiers that automatically categorize news stories from different providers into a set of news categories.

In addition to support for full-text search of news articles from multiple providers, news aggregation services such as *Google News* provide value by identifying topics that are generally considered important by a large number of news providers. In addition, aggregators allow users to compare coverage of a story between different providers, leading to a greater variety of perspectives than one single organization can offer.

In addition to *Google News,* there are a variety of aggregation services that support similar features. For example, *Topix* (*http://www.topix.net*) is similar to *Google News*, but emphasizes news categorization and geo-coding. The service automatically labels stories with the location of events, and also categorizes stories based on detected entities such as company names, celebrities or sports teams to enable fine-grained story categorization.

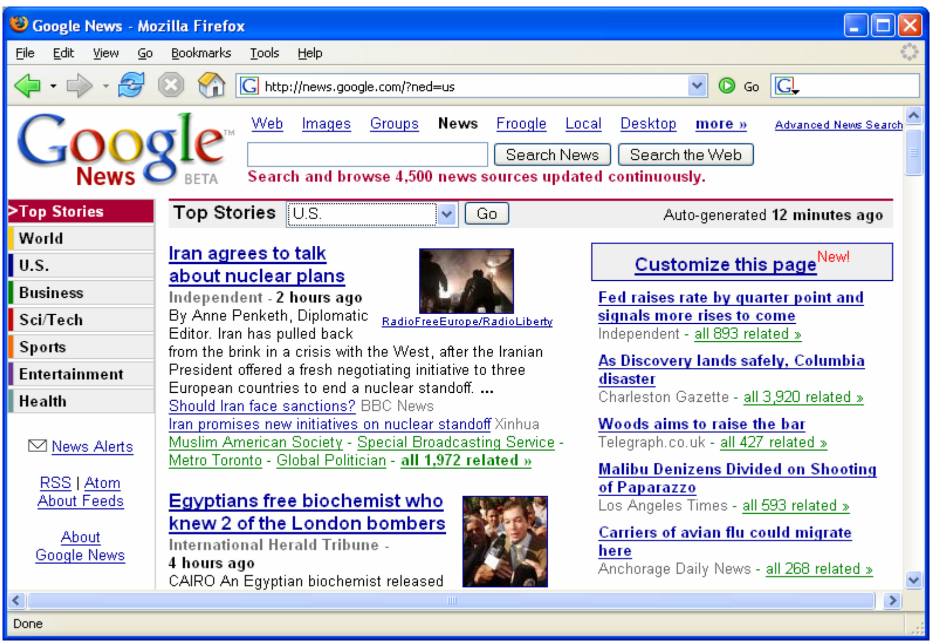

Fig. 18.3. Google News' automatically generated front page. The service aggregates news content from 4,500 sources.

18.3 Case Study

In this section, we present a case study of a personalized news service that was first released in 1999, powering the mobile version of various publicly available news services for several years [3], [4], [5]. We first describe the goals, design and user interface of the system. Next, we introduce the system's personalization algorithm – a machine learning approach specifically designed for adaptive news access. Finally, we summarize results from an empirical evaluation of the service (for additional details, see [3]).

While the study presented in this section focuses on the design and evaluation of only one individual system, numerous other studies that explore the design and utility of adaptive access to news, usenet or blog content can be found in the literature (see Section 18.2 and [14], [23], [22], [34], [11], [21], [16], [32], [2], [20]).

18.3.1 Adaptive News Personalization for Mobile Content Access

While personalization has proved to be an important supplement to web applications, the constraints of mobile information access make personalization essential to producing usable applications. Mobile devices, such as cell phones or PDAs (personal digital assistants), have much smaller screens, more limited input capabilities, slower and less reliable network connections, less memory and less processing power than desktop computers. In this section, we briefly summarize the main features of an adaptive news service for mobile content delivery that automatically infers the user's interests based on previously accessed content and personalizes content accordingly.

The news system dynamically generates a user interface that can be rendered on PDAs and cell phones. The interface displays section names (such as 'Sports'), headlines and articles. It is intended to be used by a single news site to deliver its content to readers of that site, rather than aggregating news across multiple sites. Personalization reorders sections of the news site so that the most frequently accessed sections may be reached without scrolling, reorders headlines within a section so that the most personally relevant items are displayed toward the top, and selects headlines for display on the front page.

Figure 18.18.4 illustrates how the system adapts to two different users. Both are shown the same three headlines initially. On the top row, a user reads a college football story and when the next page of headlines is requested, additional college football stories are shown. In the bottom row, a user instead reads a horse-racing story and is shown additional stories on this topic. In each case, a golf story is included on the next page, both to allow some diversity in the stories and to present the system with the opportunity to learn more about the user.

18.3.2 Learning User Models for News Access

The server that powers the described system uses a machine learning approach to automatically learn a simple model of each user's individual interests. The algorithm is a content-based approach specifically designed for news access, and addresses most of the news-specific personalization issues identified in Section 18.2.1. In short, a combination of similarity-based methods, e.g. [10], and Bayesian methods, e.g. [11], achieves the right balance of learning and adapting quickly to changing interests while avoiding brittleness. These two algorithms form a multi-strategy learning approach that learns two separate user-models: one represents the user's short-term interests, the other represents the user's long-term interests. Distinguishing between short-term and long-term models has several desirable qualities in domains with temporal characteristics [9]. Learning a short-term model from only the most recent observations may lead to user models that can adjust more rapidly to the user's changing interests. The need for two separate models can be further substantiated by the specific task at hand, i.e. classifying news stories. Users typically want to track different "threads" of ongoing recent events - a task that requires short-term information about recent events. For example, if a user has indicated interest in a story about a current Space Shuttle mission, the system should be able to identify follow-up stories and present them to the user during the following days. In addition, users have general news preferences, and modeling these general preferences may prove useful for deciding if a new story, which is not related to a recent rated event, would interest the user. With respect to the Space Shuttle example, we can identify some of the characteristic terminology used in the story and interpret it as evidence for the user's general interest in technology and science related stories.

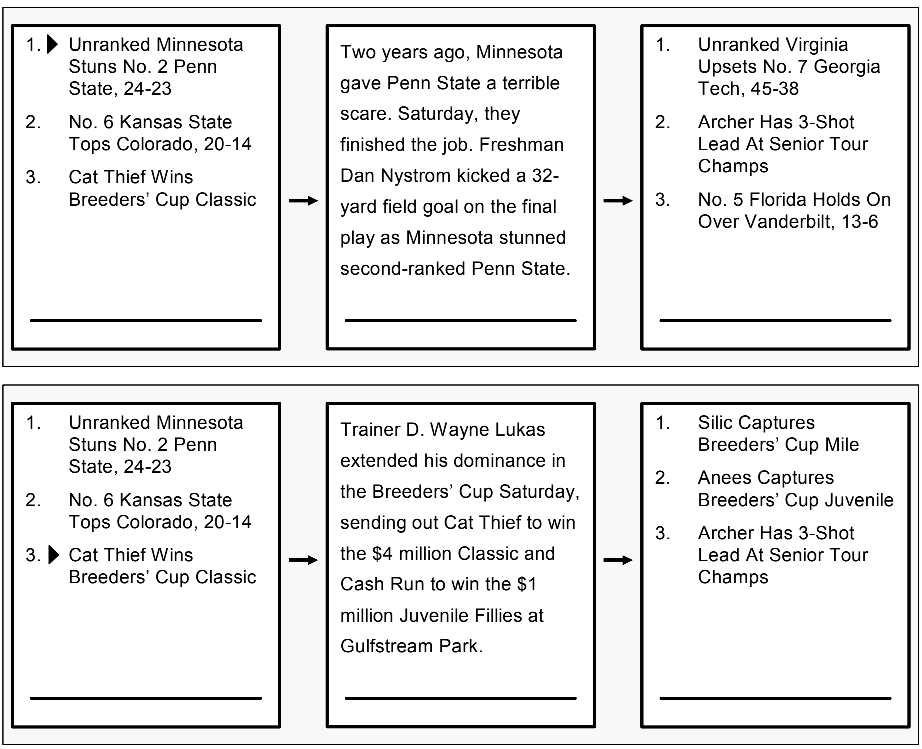

Fig. 18.4. Two different user interactions illustrating the effects of adaptive personalization. The top-row user is interested in college football, the bottom-row user is interested in horse racing. The system automatically adapts to the users' respective interests.

The purpose of the short-term model is two-fold. First, it contains recently read stories, so that other stories which belong to the same event thread can be identified. Second, it allows for identification of stories that the user already knows. A natural choice to achieve the desired functionality is the *k*-nearest-neighbor algorithm (Chapter 10 of this book [27]). To apply the algorithm to news stories, standard IR techniques, such as *tf-idf* term vectors and the cosine similarity measure are used ([33], Chapter 5 of this book [25]): news stories that were accessed or skipped by the user are represented as term vectors that are labeled as "interesting" or "not interesting", and the resulting set of stories is used to classify previously unseen news content. The model size is limited to the *n* most recent stories, so that the model remains dynamic and only reflects the user's most recent interests. To make sure that the user does not repeatedly see stories that are virtually identical to previously read content, the system artificially reduces the scores of stories that exceed a specified similarity threshold to at least one of the accessed stories in the user model. If a story to be classified does not have any near neighbors, the story cannot be classified by the short-term model at all, and is passed on to the long-term model. The nearest-neighbor-based short-term model is a reasonable choice for news recommendation, because it is able to represent a user's multiple interests, and can quickly adapt to new or changing interests. For example, a single story of a new topic is enough to allow the algorithm to identify future follow-up stories.

The system's long-term model is intended to model a user's general preferences. Since most of the words appearing in news stories are not useful for this purpose, the system periodically selects an appropriate vocabulary for each individual news category from a large sample of stories. After feature selection, the same set of features is used for all users. The goal of the feature selection process is to select informative words that recur over a long period of time. In this context, an informative word is one that distinguishes documents from one another, and can thus serve as a good topic indicator. With respect to individual documents, *tf-idf* weights can be interpreted as a measure of the amount of information that an individual word contributes to the overall content of a document. In order to determine the *n* most informative words for each document, the system sorts words with respect to their *tf-idf* values and selects the *n* highest-scoring words. A word is a useful feature for the long-term model if it frequently appears in top *n* lists over a large set of documents from one category (the system uses the most recent 10,000 news stories per category for feature selection). The system sorts all words that appear in the overall vocabulary with respect to the number of times they appear in top *n* lists. Finally, the *k* most frequent words are selected. This approach performs well at selecting the desired vocabulary: it selects words that occur frequently throughout one news category, but are still informative as measured by their *tf-idf* weights. For example, the following list shows the top 50 long-term features selected from a set of 10,000 science news stories:

```
drug, cancer, space, cells, patients, women, crops,
gene, launched, disease, food, virus,  rocket, city,
mission, bacteria, infection, children, heart, hiv,
satellite, eclipse, blood,  genetic, suns, winds,
trial, mice, orbit, antibiotics, vaccine, resistance,
russian, human,  aides, storm, percent, brain, fda,
cdc, mosquitoes, energy, test, damage, hurricane,
computer, baby, government, hospital, texas.
```

The selected features are used as part of a probabilistic learning algorithm, a naïve Bayesian classifier ([11], Chapter 10 of this book [27]), to assess the probability of stories being interesting, given that they contain a specific subset of features. Each story is represented as feature-value pairs, where features are the words from the selected feature set that appear in the story, and feature values are the corresponding word frequencies. In order to take advantage of the word frequency information, the system uses a multinomial version of naïve Bayes [24].

To predict whether a user would be interested in a news story, the system applies the two models sequentially. It uses the short-term model first, because it is based on the most recent observations only, allows the user to track news threads that have previously been rated, and can label stories as already known. If a story cannot be classified with the short-term model, the long-term model is used. In addition to the user model's prediction, editorial input is incorporated by boosting the priority of lead stories. The effect of this boosting is that first-time users of the wireless news site see articles in default order (determined by an editor), and all users always see the lead story in each section. This also allows the adaptive personalization engine to learn more about each user. Finally, the similarity-based methods are also used to ensure that a variety of news articles are presented on each screen in much the same way that a newspaper does not fill the front page with articles on the same topic.

18.3.3 Evaluation

The personalization server reorders news stories with respect to users' individual interests. The main intuition is that such a modified order helps users access relevant content. However, information is rarely presented in random order. For example, editors prioritize news stories based on human judgment, which means that, in this case, users access content in an order deemed appropriate by human professionals. While such an order is static in the sense that it is the same for every user, it is possible that it is sufficient for most users to easily access relevant content. In this section, we briefly summarize the results from two experiments that compare personalized information access provided by the described personalization server to static information access. These results show that the system's user modeling algorithm generates an adaptive order that has two closely related effects: it simplifies locating relevant content and leads to an overall increase in accessed information. The main idea underlying both experiments is to present items either in static or adaptive order so that resulting differences in users' selection and browsing behavior can be quantified. The experiments described in the following sections were performed as part of a publicly deployed news service for handheld devices, such as cell phones and PDAs with wireless Internet access [3]. Since the experiments did not require any user interface changes, all collected data is based on normal system usage by regular users who did not know that their news access patterns contributed to the system's evaluation.

The system's ultimate goal is to simplify access to interesting content. A simple and informative measure that quantifies progress towards this goal is the average display rank of selected stories. If the system successfully learned to order items with respect to users' individual interests, this would, on average, result in interesting stories moving toward the top of users' personalized lists of items. Therefore, the average display rank quantifies the system's ability to recommend interesting news stories. Since this measure does not depend on a predicted numeric score or classifying news stories based on predicted interest levels, it is possible to apply it to static information access, allowing for a comparison of both strategies. Both experiments quantify the system's personalization performance using this measure.

The "Alternating Sessions" Experiment. The "alternating sessions" experiment quantifies the difference between static and adaptive information access by randomly determining whether a user receives content in static or adaptive order. During a period of two weeks, the server used its user modeling approach for approximately half of the users, while the other half received news stories in static order determined by an editor at the news source. On odd days, users with odd account registration numbers received news in personalized order and even users received a static order. On even days, this policy was reversed. To quantify the difference between the two approaches, the server recorded the mean rank of all selected stories for the personalized and static operating modes. Since a difference between static and adaptive access can only be determined for users that previously retrieved several stories, the analysis was restricted to users with a minimum of five selected stories. Comparing both access modes for this subset of users revealed a significant difference. The average display rank of selected stories was 6.7 in the static mode and 4.2 in the adaptive mode (based on 50 users that selected 340 stories out of 1882 headlines). The practical implications of this difference become apparent by analyzing the distribution of selected stories over separate headline screens (every screen contains 4 stories). Figure 18.5 illustrates these two distributions. In the static mode, 68.7% of the selected stories were on the top two headline screens, while this was true for 86.7% of the stories in the personalized mode. It is reasonable to argue that this makes a noticeable difference when working with handheld devices. In addition, this result suggests that effective personalization can be achieved without requiring any extra effort from the user.

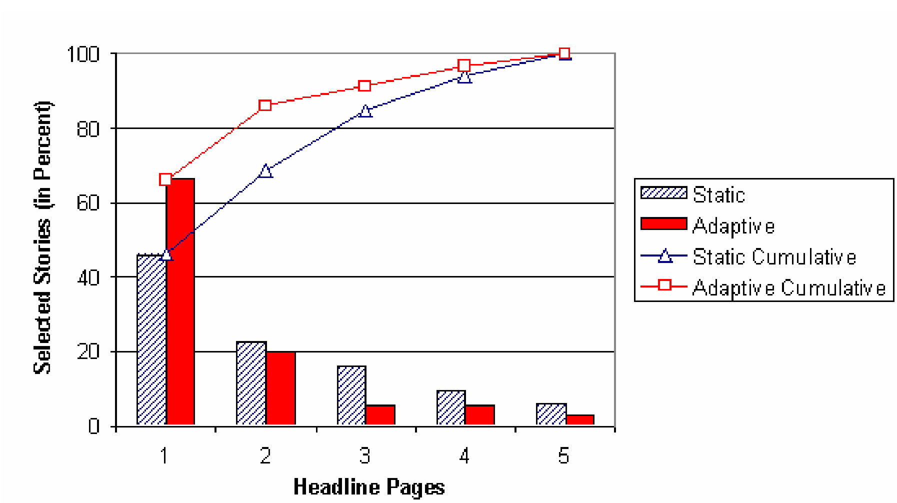

Fig. 18.5. Distribution of selected stories (alternating sessions experiment)

While the results from the first experiment look promising, the experiment has several shortcomings. First, the results are based on a small data set, consisting of only 50 users who selected 340 stories (since users can likely perceive the effect of disabling personalization, data collection was limited to two weeks). Second, due to the high cost of information access on wireless devices (some cell phone plans treat data access like regular voice minutes) users typically only select a small number of headline screens in each session. It is likely that users select from these few screens the stories that interest them most, and that this is true for both the static and adaptive access modes. Therefore, a drawback of the "alternating sessions experiment" is that users might not see stories they would have seen in the adaptive mode. Likewise, in the adaptive mode, users might not see stories they would have seen in static mode. The following experiment addresses this problem by displaying both adaptive and static stories on the same screen.

The "Alternating Stories" Experiment. The "alternating stories" experiment is similar in principle to the "alternating sessions" experiment, i.e. it is designed to quantify the difference between static and adaptive information access. However, the "alternating stories experiment" displays stories selected with respect to both the adaptive and static strategies on the same screen. During the experiment, the client was configured to display four stories on each screen, with every screen containing two adaptive stories and two static stories. The server determines randomly if the first displayed story is a static or adaptive story, and the remaining stories are selected by alternating between the two strategies. The "alternating stories" methodology has two advantages. First, the system still adapts to the users' interests, because every screen contains two stories that were selected adaptively. This results in a change of system behavior that is much more subtle from a user perspective than the resulting change of the "alternating sessions" experiment. Therefore, it is possible to run the experiment over a longer period of time, because all users still receive a useful service. Second, users see the current top-ranked adaptive and static stories on the same screen, allowing for a direct comparison between the two selection strategies. If the system learns to adjust to users' individual interests, users can be expected to select more adaptive stories when presented with a choice between adaptive and static content.

The personalization server used the "alternating stories" methodology over a period of four weeks to collect access data for 5000 adaptive stories and 5000 static stories that were shown to users who had previously selected a minimum of 5 stories. Using these criteria, data obtained from 222 different users were included in the experiment. Similar to the "alternating sessions" experiment, the average display rank can be used to quantify the difference between the two display strategies. However, using the "alternating stories" methodology, the difference between the two average display ranks was not as pronounced as in the "alternating sessions" experiment: 5.8 for the static mode vs 5.27 for the adaptive mode. Likewise, the distributions of selected stories over consecutive headline pages revealed only a small difference between the two display modes: for the static mode, 75.57% of the selected stories were on the top two headline screens, while this was true for 80.44% of the stories in the adaptive mode. The smaller difference between the two modes can be attributed to the presence of adaptive stories on every page. As a result, the user's information need might be satisfied after seeing only a small number of headline pages. If users do not have to request multiple screens to find relevant information, the observable difference in display ranks is reduced. However, this explanation only holds if users indeed select more adaptive stories than static stories. The percentage of selected stories for the two display modes clearly indicates that users are more likely to select adaptive stories than static stories. In particular, users selected 13.26% of all displayed static stories (663 stories), vs 19.02% (951 stories) of all displayed adaptive stories, which amounts to a 43% increase in selected content. Figure 18.6 shows how this difference is distributed over consecutive news headline screens. For each headline screen, this plot compares the probability that a selected story was an adaptive story to the probability that the story was presented in static order. More formally, the plot compares the conditional probabilities *p(adaptive | selected)* and *p(static | selected)* for separate headline screens.

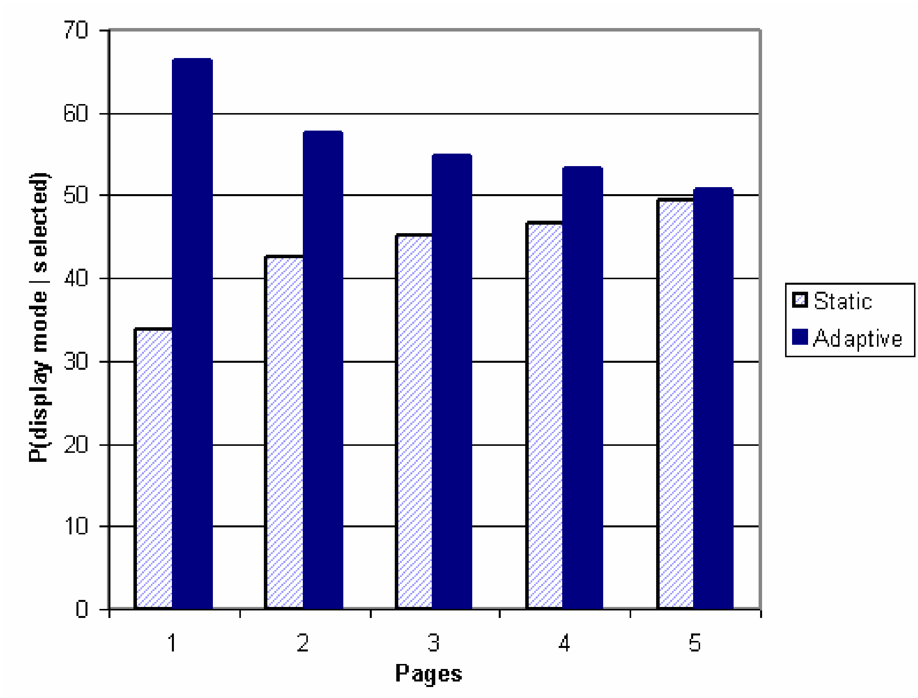

Fig. 18.6. Static vs adaptive selection probabilities (alternating stories experiment)

Figure 18.6 also shows that the difference in selection probabilities is particularly noticeable on the first headline screen and then decreases gradually from page to page. On the first headline screen, *p(static | selected)* is 0.33 vs 0.66 for *p(adaptive | selected)*. This difference indicates that the adaptive display strategy indeed helps users locate relevant content, as users prefer adaptive stories over static stories on average.

In summary, the “alternating sessions” and “alternating stories” experiments both show that adaptive information access is superior to static access. The “alternating sessions” experiment demonstrated that the adaptive order helps to move interesting items towards the beginning of personalized item lists, simplifying access to relevant content. The “alternating stories” experiment showed that the system is capable of ordering content in a way such that the top-ranked stories have a significantly higher chance of being selected than the top-ranked stories obtained from a static order.

18.4 Recent Trends and Systems

In this section, we briefly discuss recent news delivery trends, novel systems and emerging research opportunities.

18.4.1 Personalizing Audio and Video News Feeds

The Internet is rapidly turning into an on-demand delivery platform for multimedia content. For example, as part of the phenomenal success of *Apple Inc’s* portable audio player, the *iPod,* online audio distribution of news content, or in short *podcasting,* is becoming increasingly popular. Thousands of regularly updated news programs can be located via countless services (including *Apple Inc*’s popular *iTunes* software) and downloaded to personal media players. Even though podcasting is still in its infancy, selecting the most informative, relevant or interesting audio content from thousands of feeds is challenging. Since information is distributed in the form of audio files, users currently cannot search for information within podcasts, and the text-based recommendation techniques discussed in this book cannot be used (because textual transcripts are usually not available for audio broadcasts). Since collaborative filtering algorithms only depend on ratings from other people and do not require any content analysis, these techniques are immediately applicable to audio feed recommendation. However, as discussed in Section 18.2.1, collaborative filtering has potential disadvantages in the context of the very dynamic nature of news content. Therefore, content-based recommendations for audio files, or fragments thereof, could significantly enhance the utility of available audio content. For example, users may not want to listen to an entire audio program if only a small segment of it is personally relevant or interesting. Techniques that automatically extract text from audio files, enable full-text search, find topic-based segments within audio or use content-based recommendation techniques to assemble personalized podcasts, is an interesting area for future work. Clearly, video feeds face similar problems: textual transcripts are often not available, and features that could be automatically extracted by image analysis techniques are not yet semantically meaningful enough to support accurate news personalization. However, since this is an area that will undoubtedly become increasingly important, adaptive access to multimedia content is an active area of research [22], [34], [15], [5].

18.4.2 Personalization and the Blogosphere

The term *blogosphere* refers to the collective set of all weblogs or *blogs.* As *blogging* becomes an increasingly popular form of self-expression, there is a need for more sophisticated tools to help navigate the blogosphere and discover relevant content. Initial systems that support personalized blog access are beginning to emerge: e.g. *Findory.com,* as described in Section 18.2.1, can adaptively personalize blogs, similar to its news personalization capabilities. *NewsGator (www.newsgator.com)* can recommend news and blog feeds, based on a collaborative approach that uses a user's subscriptions as the basis for personalized feed recommendations. In addition, blog search engines, such as *www.technorati.com* or *www.blogdigger.com* are beginning to incorporate customization features (similar to most news services) into their sites. However, these customization features are usually static and do not adapt to the user's interests.

18.4.1 News Zeitgeist

Zeitgeist is a German word that means "the spirit *(Geist)* of the time *(Zeit)*". As news and blog aggregation services are becoming more popular, many sites are incorporating *Zeitgeist* features. The goal is to automatically identify the most popular or talked-about topics, as an expression of the current blogosphere "buzz" or *Zeitgeist.* For example, *www.technorati.com* lists the most popular blog searches of the hour, the most talked about books and movies, and most-cited blogs. *Daypop (www.daypop.com)* uses blogs as a *Zeitgeist* meter for news content, by generating a list of the 40 news stories that are most frequently cited in the blogosphere. Closely related to this, *Digg (www.digg.com)* is a representative of a new breed of social content discovery sites: users submit potentially interesting news articles or blog posts to the site, and *Digg's* user community expresses interest in the submitted stories by clicking corresponding "digg it" buttons. The stories with the most "diggs" are then prominently displayed on the site, which means that, arguably, the content featured on *Digg* automatically adapts to the interests of its user base.

In summary, as news and blog services evolve, we will continue to see meta-services that not only aggregate content, but also attempt to automatically convey *Zeitgeist* information – by auto-generating pages that adapt to the spirit of the time.

18.5 Summary and Conclusions

The web is increasingly evolving into the most powerful news delivery platform of the 21st century. It provides new information dissemination channels for established news organizations, allows individuals to be heard, and enables new forms of coverage analysis. As our established reading, publishing and sense-making practices continue to evolve, we need new technology to help leverage the full potential of web-based news distribution. Continued growth of online news content is of limited use if we cannot find the personally most relevant and useful information. This section outlined early steps towards technology that is specifically designed to help us navigate the continuously evolving news landscape.

References

1. Balabanovic, M.: Learning to Surf: Multiagent Systems for Adaptive Web Page Recommendation. Ph.D. Thesis, Stanford University (1998)
2. Bharat, K., Kamba, T., Albers, M.: Personalized Interactive News on the Web. Multimedia Systems 6(5) (1998) 349-358

3. Billsus, D., Pazzani, M.: User Modeling for Adaptive News Access. User Modeling and User-Adapted Interaction 10(2/3) (2000) 147-180
4. Billsus D., Pazzani, M., Chen, J.: A learning agent for wireless news access. In: Proceedings of the 2000 International Conference on Intelligent User Interfaces (2000) 33-36
5. Billsus, D., Brunk, C., Evans, C., Gladish, B., Pazzani, M.: Adaptive Interfaces for Ubiquitous Web Access. Communications of the ACM 45(5) (2002) 34-38
6. Billsus, D., Hilbert D., Maynes-Aminzade, D.: Improving Proactive Information Systems. In: Proceedings of the 10th International Conference on Intelligent User Interfaces (2005) 159-166
7. Borthwick, A.: A Maximum Entropy Approach to Named Entity Recognition. Ph.D. Thesis. New York University (1999)
8. Budzik, J., Hammond K., Birnbaum, L.: Information Access in Context. Knowledge-Based Systems 14 (1-2) (2005) 37-53
9. Chiu, B., Webb, G.: Using decision trees for agent modeling: improving prediction performance. User Modeling and User-Adapted Interaction, 8 (2005) 131–152
10. Cover, T., Hart, P.: Nearest Neighbor pattern classification, IEEE Transactions on Information Theory, 13 (1967) 21-27
11. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. New York, NY: Wiley and Sons (1973)
12. Gabrilovich, E., Dumais, S., Horvitz, E.: Newsjunkie: Providing Personalized News Feeds via Analysis of Information Novelty. In: The Thirteenth International World Wide Web Conference (2004) 482-490
13. Gauch, S., Speretta, M, Chandramouli, A., Micarelli, A.: User Profiles for Personalized Information Access. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York (2007) this volume
14. Haas, N., Bolle, R., Dimitrova, N., Janevski, A., Zimmerman. J.: Personalized News through Content Augmentation and Profiling. In: IEEE 2002 International Conference on Image Processing, (ICIP 2002) (2002) 9-12
15. Haggerty, A., White, R.W. and Jose, J.M.: NewsFlash: Adaptive TV News Delivery on the Web. In: Proceedings of the 1st International Workshop on Adaptive Multimedia Retrieval (2003) 72-86
16. Jokela, S., Turpeinen, M., Kurki, T., Savia, E., Sulonen, R.: The Role of Structured Content in a Personalized News Service. In: Hawaii International Conference on System Sciences (HICSS 34) (2001) 1-10
17. Kelly, D., Belkin, N.: Reading Time, Scrolling and Interaction: Exploring Sources of User Preferences for Relevance Feedback During Interactive Information Retrieval. In: Proceedings of the 24th Annual ACM SIGIR Conference (2001) 408-409
18. Kelly, D., Teevan, J.: Implicit Feedback for Inferring User Preference: A Bibliography. SIGIR Forum 37(2) (2003) 18-28
19. Klinkenberg, R., Renz, I.: Adaptive Information Filtering: Learning Drifting Concepts. In: AAAI/ICML-98 Workshop on Learning for Text Categorization, Technical Report WS-98-05, AAAI Press (1998) 33-40
20. Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., Riedl, J.: GroupLens: Applying Collaborative Filtering to Usenet News. Communic. of the ACM 40(3) (1997) 77-87
21. Lai, H., Liang, T., Fu, Y.: Customized Internet News Services Based on Customer Profiles. In: Proceedings of the 5th International Conference on Electronic Commerce (ICEC 2003), Pittsburgh , Pennsylvania (2003) 225-229
22. Lee, H., Smeaton, F., O'Connor, N., Smyth, B.: User Evaluation of Físchlár-News: An Automatic Broadcast News Delivery System. ACM Transactions on Information Systems 24(2) (2006) 145-189

23. Magnini, B., Strapparava, C.: User Modeling for News Web Sites with Word Sense Based Techniques. User Modeling and User Adapted Interaction 14 (2004) 239-257
24. McCallum, A., Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. In: AAAI/ICML-98 Workshop on Learning for Text Categorization, Technical Report WS-98-05, AAAI Press (1998) 41-48
25. Micarelli, A., Sciarrone, F., Marinilli, M.: Document Modeling. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York (2007) this volume
26. Mobasher, B.: Data Mining for Web Personalization. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York (2007) this volume
27. Pazzani, M., Billsus, D.: Content-based Recommendation Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York (2007) this volume
28. Rhodes, B., Maes, P.: Just-in-Time Information Retrieval Agents. IBM Systems Journal 39 (2000) 685-704.
29. Sakagami, H., Kamba, T.: Learning Personal Preferences on Online Newspaper Articles from User Behaviors. In: Proceedings of the Sixth International World Wide Web Conference (WWW6) (1997) 291-300
30. Schafer, B., Frankowski, D., Herlocker, J., Sen, S.: Collaborative Filtering Recommender Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York (2007) this volume
31. Smyth, B., Cotter, P.: Intelligent Navigation for Mobile Internet Portals. In: IJCAI Workshop on AI Moves to IA: Workshop on Artificial Intelligence, Information Access, and Mobile Computing. The 18th International Joint Conference on Artificial Intelligence (IJCAI-03) http://www.dimi.uniud.it/workshop/ai2ia/ (2003)
32. Sorensen, H., McElligott, M.: PSUN: A Profiling System for Usenet News. In: Proceedings of the CIKM 95 Workshop on Intelligent Information Agents. http://www.cs.umbc.edu /cikm/iia/proc.html (1995)
33. Van Rijsbergen, C. J.: Information Retrieval. London: Butterworths (1979)
34. Wang, Q., Balke, W-T., Kieling, W., Huhn, A.: P-News: Deeply Personalized News Dissemination for MPEG-7 Based Digital Libraries. In: Research and Advanced Technology for Digital Libraries: 8th European Conference (2004) 256-268
35. Wayne, C.: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. In: Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece (2000) 1487-1493
36. Widyantoro, D., Ioerger, T., Yu, J.: An Adaptive Algorithm for Learning Changes in User Interests. In: Proceedings of the Eighth International Conference on Information and Knowledge Management (1999) 405-412