Toward An Adaptive Web: The State of the Art and Science

M. Kilfoil, A. Ghorbani, W. Xing, Z. Lei, J. Lu, J. Zhang and X. Xu

Faculty of Computer Science
University of New Brunswick
Fredericton, NB, E3B 5A3, Canada

ABSTRACT

As the World Wide Web matures, it rapidly grows in both size and complexity. In this expanding environment, the needs and interests of individual users become buried under the sheer weight of possible viewing choices. To counter this, there has been a rise in research in *adaptive websites*, a combination of data mining, machine learning, user modeling, Human Computer Interaction (HCI), optimization theory and graph theory which seeks to sift through the tides of possible pages to provide users with a 'high-quality' stream of information.

This paper provides a description of adaptive website research, including the goals aimed at, the challenges discovered and the approaches to solutions.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation (e.g., HCI)**]: Hypertext/Hypermedia; I.2.6 [**Artificial Intelligence**]: Learning—*knowledge acquisition*

General Terms

Human Factors, Performance, Measurement

Keywords

Adaptive Hypermedia, Adaptive Websites, Personalization, Recommender Systems, Authoring, User model

1. INTRODUCTION

Websites are becoming more and more popular and convenient for providing information over a broad number of topics. They are used in diverse systems, such as educational systems (for example, the ELM-ART II system [53]), online information systems, on-line help systems, information retrieval systems, and e-commerce systems [9]. Problems occur with this increased use of websites, including:

1. The rich link structure of a hypermedia application can cause users to get easily overwhelmed by the sheer number of navigation choices, and they may become unable to navigate effectively; this is referred to as the "lost-in-hyperspace" problem of navigating the Web.

2. Websites provide (from the point of view of a user) relatively static content, but they are viewed by diverse users. This may cause difficulty for those who have less background, may be redundant for those who already know the information, or may present more uninteresting than interesting material for others. This is referred to as the 'one-size-fits-all' problem of non-adaptive websites.

In order to cope with an increasingly large and complex World Wide Web, there is a demand for intelligent tools and structures which can simplify the experience and make navigation of the sites easier for users and yet maximize the quality and completeness of the experience. These tools and structures should provide sufficient intelligence so that one can sense the environment, perceive and interpret the situations in order to make decisions and to control actions. From our point of view, this is possible through the integration of techniques from multiple disciplines and research areas.

One reasonable approach to reduce such difficulties is to combine artificial intelligence, user modeling, graph theory, and information mining techniques to create websites and website browsing tools which are adaptive. *Adaptive* refers to the ability of the website or tool to change its behaviour or responses in reaction to the way it is used, or the needs of its users.

The broadest definition of an adaptive website is a *website which changes based on the way it is used* [46]. Changes can take on many forms (as described in section 3), may either be immediate (as in the case of recommendation systems) or gradual (as in the case of systems which suggest changes to a website administrator). An adaptive website technique may apply to smaller, closed-corpus websites, where the entire site is known in advance, or to the Web in general, where it is virtually impossible to know all the web pages even casually[1]. In the latter case, we might describe it as an

[1]One coverage estimate in 1998 for public search engines [35] suggested that no more than approximately a third of all of the Web was indexed by any one search engine. With the rapid turnover and continual exponential growth of the Web, this seems unlikely to improve.

CNSR 2003 Conference, May 15-16, 2003, Moncton, New Brunswick, Canada.

adaptive web browsing tool rather than an adaptive website.

An adaptive website should have the ability to recognize users and events, to reason about, and plan for the future. Web logs are the main source of user behaviour data used to continuously tune and adapt the site to its users. Adaptation may be done by temporarily altering text, links or page format. Creating new pages and adding or removing links are also possible permanent adaptations that may be considered.

Adaptive websites can filter out and prioritize links and content based on knowledge of the user's interest and behaviours or general knowledge of all users' behaviour and interests. This simplifies the user's browsing and makes each page viewed or link followed more relevant to their needs, which improves the overall quality of their experience.

Adaptive web technologies belong properly as a subset of the field of adaptive hypermedia, but the rapid growth of the Web has magnified its importance and prevalence to the point where it currently overshadows the larger field. This paper focuses on issues and technologies specifically relating to the Web, mentioning those that belong to the larger field only for historical reasons or when there are useful abstractions or generalizations.

Section 2 describes the desired goals that adaptive web site research seeks to satisfy and the challenges that are encountered when pursuing these goals. Section 3 describes some of the ways in which adaptations may be carried out. Section 4 describes some of the current general approaches researchers are taking are given. Some of the more prominent or interesting implementations are described in Section 5. Finally, the conclusions of the study are summarized in Section 6.

2. GOALS AND CHALLENGES

2.1 Goals

In order to discuss the goals of an adaptive web systems, it is necessary to describe the kinds of users this system must satisfy. In any website, consumers of the services can usually be categorized into the following three groups:

Individual Users: These are individual users who receive personalized recommendations based on their interests and activities.

User Groups: Users who are like-minded and receive generalized recommendations based on what others of their group are interested in or actively pursuing.

Website Operators: Administrators, content owners and sponsors form a different group of users which might be satisfied by an adaptive system. This group receives analytical information about the way the website is being used, to determine if it is being used as intended.

To satisfy these users, there are a number of general goals which should to varying degrees be satisfied. These include *personalization*, *recommendation*, *selection* and *usage analysis*.

Personalization: In order to make an adaptive website more effective for individual users, it should reflect their own interests, needs, knowledge, background or goals. This process of reflection and alteration is called *personalization*. It is done through combined use of various technologies (collaborative filtering and profiling, for example) and web visitors' information to personalize interactions between a web server and each individual visitor. Personalization's primary objective is to tailor (alter) the site in order to accommodate a customer's stated needs using information either previously obtained or provided in real-time about the visitor.

Recommendation: To address the problem of users lost in the overwhelming number of possible web pages, adaptive websites may choose to provide a shortlist of *recommendations* which contains links to the most relevant, most important or most accurate pages for that user. One of the forms of recommendation is "path-shortening", where a probable eventual destination that a user will want to visit is listed as a link earlier. This attempts to make more relevant eventual links available earlier to a user, removing the need for the user to navigate through intervening pages.

Selection: This refers to the selective removal or inclusion of items from a larger set of items based on some criteria or rules. In adaptive websites, sections from a web page might be removed if they are irrelevant to a user or included if they are relevant.

Usage Analysis: It is a difficult task to determine if a website is "effective" in performing its function, such as determining if an e-commerce site leaves people lost or if it is difficult to navigate between pages with related information on a particular site. The analytical side of an adaptive website may provide a way to measure and improve this effectiveness by automatically suggesting relevant information and pages to users, measuring the way the website is used and suggesting changes to the site to administrators which should improve the users' experiences.

User's interests can be identified from the pages they visit and the amount of time they spend on them. Revisiting a certain page and spending more time on it may be considered an indication of strong interest in that page [36].

2.2 Challenges

The environment in which an adaptive web system operates presents certain challenges which impact their feasibility and performance.

2.2.1 Lack of Information

In order to decide if an adaptation should be made, the system has to have data of various kinds. Such data may include website structure, website content, user profiles, and website usage data.

In most cases, data is rather sparse, or a great deal of interpretation must be done to turn it into actually useful information.

Website structure: The way that the website is physically laid out can be useful toward understanding usage behaviour and interpreting system suggestions. Furthermore, the *semantic* information about the reasons why the structure exists in the way that it does may

also inform the system [28]. In closed-corpus systems, it is possible to know this in detail, but in Web-spanning systems this information is not immediately available, and retrieving it fully is an enormous problem of scale (and may not even be possible). Note that an additional problem occurs when one considers the changeability of web pages: the website structure is easily changed, and is subject to change over time.

Website content: The content of web pages themselves is essential to determining particular topical interests and to understand the relationships between pages [39, 26] , but suffers even more greatly from the problems of scale, availability and changeability.

User data: Information about the people using the system can help in understanding their interests, or in finding common groups of users who share interests. User profiling is a technique by which user data is gathered from one or more sources, processed and analyzed in order to better understand a user's characteristics and browsing trends. User profile data may be gathered from the client side, server side or from a proxy, either through direct interview or through observed behaviour such as purchases or dialogue acts [54]. The data may be categorized as *demographic*, *behavioural*, *attitudinal* or *click stream* data [31]. There are really two types of user data: those that describe individuals and those that describe groups of users.

Each individual user profile is based on contextual relevance observed during the user information access [37]. Information may include demographics, goals and interests, browsing behaviour patterns, browsing capabilities, shopping behaviours, connection speed and type, and human relationships.

One of the key problems with user information is the difficulty in obtaining it, and another is the difficulty of verifying the veracity of the data [6]. It has been suggested that users are neither interested in providing this information, nor are they necessarily even willing to provide it for privacy concerns [30].

Website usage data: Perhaps the most important data set is the recording of interactions of users with the website, in other words, the way that the website is used. This data set may be described in terms of simple page views, transactions (which are "significant" events, and may combine multiple page views), and sessions (which are a combination of page views or transactions that together represent an individual users' experience) [16]. In addition to the simple sequence of events, information about time of access and frequency of access is also useful.

This is by far the most abundant collection of data, provided primarily by web server logs. It is perhaps the least informative on its own; without the identification of individuals and some concept of the structure of the website, it is difficult to derive much useful information out of this data.

2.2.2 Measurement

In order to assess the effectiveness of adaptive website technologies, one has to create measures to examine the impact and quality of the adaptations performed or suggested.

While there are a large number of metrics which measure the Web (see [20] for an extensive treatment of many of them), or seek to describe user behaviour [43], these measures are far from absolute. In each case, the assumptions in the models used (see Section 4.2) greatly colour the perception of the meaningfulness and significance of the metric.

2.2.3 Impact on User Experience

When adaptation takes place, there is by definition some changes which are made to the website, perhaps to the content of the pages, the structure of the site or the links which are presented to the user.

Since the website is changing, it is important to consider the impact that making a change would have on the user's experience, and avoid or modify changes in light of how the experience should be maintained. For example, in a website which has a highly visual layout, the addition or removal of links may have a disastrous effect on that layout; even the modification of the colour of the links or the augmentation of link icons might confuse the user as to what links they had visited already as opposed to which links they have yet to visit [10]. In another case, while there may be a large number of links which are deemed relevant to a particular user, some subset of these must be chosen to avoid overwhelming the user and putting them back into the "lost-in-hyperspace" situation. In yet another example, the adaptation of content may confuse or disorient a user, as the location of familiar items may be radically altered based on the system's perceived shift in interests.

This is an outstanding problem which is rarely addressed; it is an issue falling primarily into the realm of the study of HCI, and is usually treated as a secondary problem to the determination of when to adapt.

2.2.4 Changing Interests

The basic model of a user revolves around some measure of their intent or goals, usually described in terms of the user's interests. The simplest model assumes that these interests, once understood, do not change. A single user may in fact have multiple interests rather than a single one. For example, a user may temporarily focus on a particular topic for a period of time and then revert to a less focused interest, shifting from a long-term interest to a short-term interest and then back to a long-term interest [38]. Some users may want information about a specific topic after they explore different kinds of information, so their interests become more focused and previous explorations do not indicate general interests. Some users may need wider background knowledge after they study a specific topic, moving from a specific focus to a general view. A user's interest in a particular area may also wax and wane. These and other reasons may cause changes in a user's interests, which may happen abruptly and rapidly (*concept shift*) or gradually and slowly (*concept drift*) [34]. Ideally, adaptive web systems should be able to adapt to such interest changes.

A user's interests may not simply change, but things which interested a user in the past may become interests again at some future point. So, in addition to some form of interest *forgetting*, which represents a shift over time, there is also the idea of interest *remembering*, where old interests may reappear [32].

All of these factors complicate the view of the user's interests. Most systems use a simplified view which is based

off a single session, rather than a more complicated view learned over multiple sessions, which eases the privacy concerns caused by long-term tracking of users (see Section 2.2.12), but limits the amount of evidence that a system has to learn or approximate those interests. (For more information about approaches to dealing with changing interests, see Section 4.5).

2.2.5 Indirect Users

Similar to the problem of concept shift is the problem of indirect users. In the web environment, the immediate user is the person who is directly typing in requests, viewing web pages or searching the web. That user may actually be acting on behalf of another person who might have entirely different interests. For example, on an e-commerce site, the direct user might be making purchases for a relative or friend. That other person may be called an *indirect* user.

The idea of an indirect user suggests that an interest or activity profile built for one user may actually represent more than one user. An adaptive system might partition a perceived user profile into things which represent that direct user and things which represent one (or more) indirect users to be more effective at providing suggestions or making predictions [31].

2.2.6 Authoring

In addition to the tools which analyze actions and perform adaptations there is the issue of creating tools and methodologies which assist website developers and maintainers in the creation and modification of a website which is conducive to being adaptive [56].

2.2.7 Semantically Related Information

In many cases, we are trying to deal with data which is semantically related, such as web pages which contain the same type of information, users to whom we want to provide recommendations described by their interests, or the implication of a link placed from one page to another implying some sort of similarity, authority or other relationship (such as explicit browsing path). However, semantic information processing is really in its infancy, and is generally inefficient and limited. "Pure" information theory approaches, such as TFIDF, are generally unable to work on synonymous (but not physically similar) terms. This limits the effectiveness of any system built on such techniques. (For more detail of TFIDF and other information similarity measures, see Section 4.4.)

2.2.8 Incorrect/inaccurate Modeling

Since the system is trying to draw conclusions and common features from a less-than-precise body of information, it will on occasion have considerable difficulty in reaching accurate or even marginally correct generalizations. If there is no way to inspect the decision process made in reaching conclusions and potentially correcting them, there can be disastrous results, with the system generating entirely inappropriate suggestions (as in the case of a TiVo gone wild in [58]).

2.2.9 Lack of Negative Feedback

When examining a user's browsing behaviour it is difficult to determine whether a particular page visit should be considered a positive, negative or indifferent experience for the user. Some systems introduce a feedback mechanism to allow the user to rate their experience, while others cluster the similarity of the user's transactions to differentiate between those pages to which the user was merely exploring and those which the user actually found interesting [48].

Similarly, the idea of *backtracking* suggests that "visitors will backtrack if they do not find the page where they expect it" [50]. In this case, care must be taken to interpret this appropriately: was this a case of the user exploring to find a particular target, or was this a normal browsing behaviour?

On the other hand, some approaches, such as Bayes' theorem and k-nearest neighbour algorithm, have been used to deal with only positive evidence by employing a notion of similarity or distance [48, 49].

2.2.10 Caching

To improve the throughput of website delivery, several layers of caching have been introduced into the Web browsing experience. By definition, this means that not all pages that the user sees are actually requested each time. Because of this, log information from a web server will not necessarily have a complete record of the user's browsing behaviour, and any system which hopes to analyze the records in order to deduce common behaviour patterns will have to deal with this issue in some way.

One approach to solving this problem is the path completion algorithm, which attempts to identify when pages are missing and insert these missing pages into the session [51]. In this case, it is often assumed that the link structure of the website is already known, thus the discovery of the cached pages can be made by looking for cases where explicit links between two consecutive views (or transactions) do not exist in the structure. This approach is closely related to the idea of backtracking described in the previous section [50].

2.2.11 Noise

In all statistical and machine learning systems, noise in the input data is a factor. This is no less a case with adaptive systems. Noise in this case can come from robots and spiders visiting the site generate uninteresting usage behaviours, or from inaccurate, incomplete or unjustified observations about users (including demographic information that users provide about themselves). In each case, decisions must be made about the pre-processing of the data to handle the noise, perhaps by dropping records, filling in missing data, or massaging data [16].

The caching problem above can also be considered a form of noise, where instead of random or irrelevant information being added to the log, important and useful information is "randomly" removed.

2.2.12 Privacy

Adaptive systems which capture information about users in order to build a profile about them can be viewed as an impingement on personal privacy by some users. This issue is a social one and not a technological one, but does imply that the results of a user model that describes a user or group of users should be treated carefully and not casually. Privacy laws may restrict both the content of personal user data and the methods that may be used for processing them. Furthermore, Web systems normally face customers from all over the world. In this case, the fact that different

countries have different privacy laws may need to be taken into account in user modeling [30].

3. TARGETS FOR ADAPTATION

The heart of an adaptive web system is its ability to change in response to the way it is used. This section describes the kinds of changes that such a system may perform. The content, presentation and links of a web page are closely related, so there is bound to be crossover between these categories.

3.1 Content

One of the basic modifications that might be made is to change the content of the web page, based on the model that the system has been able to deduce about the user [31, 10]. Content might be added to or removed, or it might be simply rearranged [21]. These modifications might be done to accomplish several things, including the following:

Optional explanations: Additional explanations might be presented (or removed) to complement a user's presumed background knowledge in the subject [31].

Optional detail: Additional detailed information might be added or removed to pages depending on a user's perceived interest in the topic [31].

Personalized recommendations: Particularly in the e-commerce world, recommendations for offers or products in which the user might be interested may be presented. In other websites, this would include putting links to other conceptually-related subsections that the user might find interesting [31].

Optional opportunistic hints: Hints to understanding or discovering information might be added based on the users' interests and on current circumstances [31].

Substitution of content: Depending on the perceived browser capabilities or user interests, content of one type may be replaced with equivalent content of a lesser or greater browser requirement. For example, an image of a map might be replaced with a textual description of the map for users who are visually impaired and using a text reader, or a video might be replaced with a still picture with a link to the video for a user whose actions (or preferences) indicate a low-bandwidth connection [22, 44].

3.2 Presentation

In addition to modifying the content of the page, one can also change the way it is presented in order to serve a user.

Page variants: Different versions of all possible adaptive variations may be stored in the system, and the particular page selected at run time [31]. One common case where this is done is for multi-lingual websites, where a version for each web page translated into each language is stored, and then selected based on the user's language preference.

Fragment variants: Similar to the technique of page variants is this technique, this technique stores content fragments (or *atoms*) and selects the appropriate fragments at runtime, assembling them into a static page when needed [31]. This technique can readily be seen for any site which has easily separable atoms of content, such as news sites [3].

Fragment colouring: This technique colours fragments to highlight which ones are important and de-emphasize those which are irrelevant. In this case, the content of the pages is the same for all users; this avoids the problem of an incorrect characterization of a user having too negative an impact on their experience [31].

Adaptive natural-language generation: Similar to the example of page variants where multiple copies of a page are stored and retrieved when needed, this generates alternative text descriptions for different users [31]. A similar technique can be seen in online page translators such as Altavista's Babel Fish [1].

3.3 Links

Adaptation of navigation realizes adaptation by changing the links of the system [31, 10]. This adaptation speeds up the search for a particular page and helps to avoid the problem of users lost in hyperspace. There are several techniques to realize adaptation of navigation such as:

Direct guidance: This technique provides a user with a dynamic link, usually a "next" button, linking to the node which the system predicts to be the best for the user. It is the simplest technique for realizing adaptation of navigation [9, 10].

Link sorting: This technique first selects the most relevant pages based on the users' interests or goals, then sorts them based on their relevance, finally presenting them to the users as an ordered list of hypertext links. The most relevant link is always presented first, but if the user is not happy with this link for some reason, he or she can try the second and the following suggested links [31, 10]. This technology has two problems: it is hard to use for indexes and content pages, and it cannot be used with non-contextual links and maps. The order of links may also change frequently as the user visits pages, possibly contributing to a user's disorientation [9].

Link hiding: This technique hides links that are not relevant to the users' interests or goals [31, 10]. The links are hidden by making them look exactly like the surrounding text. This produces the appearance of less links and less confusion, and should speed up navigation. Furthermore, this technique can be used with all kinds of links indicating non-contextual, contextual, index and map links; it is more transparent to the user and is more "stable" than adaptive link sorting [9]. The real structure of the system does not change because the link still exists even though the user cannot easily find it. This change is only temporary, lasting only for that page view.

Link removal: This technique removes the link entirely from the document, leaving behind only the plain anchor text or image [19]. With this technique, the link is not available, and the user cannot follow it.

Link disabling: This technique disables irrelevant links [31]. This is similar to adaptive link hiding, but this technique removes the link but leaves the visual appearance of the link nearly untouched. This allows the removal of links which are presented inline in paragraphs, for example, without disrupting the text itself.

Link annotation: This technique uses different symbols, such as icons, colours, font sizes, or font types, to indicate relevant extent of the links [53, 9, 27, 10]. This technique can easily simulate link hiding and link removal by using special fonts, font styles, or colours [52], instead of hiding or removing irrelevant links. Moreover, this technique can avoid the "lost-in-hyperspace" problem caused by link hiding and link removal [9].

3.4 Structure

It is possible for an adaptive system to modify the long-term structure of the website in a "permanent" fashion, rather than the per-request temporary fashion suggested above. Usually, the final decision to add or remove a page or atom should ultimately be made by some human administrator, but the indication of whether it should be added or dropped can be made by the system. In this way, the adaptive system can be viewed as a tool to help the administrator measure the effectiveness of a website.

Several indications may be given by the system, including:

New index pages: Based on the perceived common viewing patterns of a group of users, the system might suggest new index pages which capture links which serve as a central point to support that group [47].

Measurement of use of a set of pages: By generating statistics about commonly viewed pages and subsets of pages, the administrator will be more informed about whether the viewing pattern matches their expectations. Pages which should be included in some groups might actually be omitted, indicating that those pages are incorrectly promoted or linked, for example.

Permanent new link suggestions: The system might suggest that certain links between pages be made permanent for similar reasons to the suggestion above that they be added for individual page views.

While the adaptation of links might also be seen as the adaptation of the structure of a website, such adaptations are of a short-term time period and have little lasting impact on the website beyond an individual browsing session. Also, normal, short-term adaptations cannot change the form and structure of image maps, which would require a human administrator to accomplish [9].

4. APPROACHES AND TECHNOLOGIES

This is a cross-section of some of the most common or promising techniques. It is meant to be illustrative, not exhaustive.

4.1 Development and Authoring Support

There are, at present, few attempts to produce standards, as the problems are not yet well solved. One such approach is the AHAM model [55, 56], which lays out the common components of a system (described above in the next section). Cooley et al [16] also describes the sequence of steps involved in general in performing adaptations. The steps are, roughly, as follows:

Data capturing: This is an ongoing activity, but most systems suggest that the bulk of this data is collected in large batches, rather than bit-by-bit. At this step, there are issues of what data to capture, how the data may be captured, and storage of the captured data to decide upon, as well as some issues of handling noise, accuracy and privacy.

Data cleaning: Also called *pre-processing*, this is possibly the most important step, as it attempts to transform the somewhat mushy log data into stronger, more complete, more usable information. It also will likely involve cleansing the data to protect privacy as well as drawing in additional information (such as web site data and permanent user data) to make the raw captured data more informative.

Discovering users, sessions and transactions: This is the identification and separation of the stream of data into discrete users, sessions of activity for those users, and "significant" transactions for each session. This is sometimes described as a part of the previous step, but it is worth mentioning on its own because it usually requires partial analysis of the data with respect to previously analyzed data (such as previous activity, other user, session or transaction identification decisions), as well as path completion.

Discovering patterns and clusters: At this stage, the activities and attributes of all of the users and pages are mined to discover any patterns or clusters of behaviour, information, or interests that might be good indicators of future behaviour, desired information or potential interests. This is the bulk of the data processing, and is generally performed offline.

Categorizing an individual user: Similar to the previous step, this step seeks to discover patterns and clusters of behaviour for an individual user in an online fashion, while the user is active. Once this information has been discovered, adaptations to aid that specific user can be performed. Some systems do not seek to improve individual active user experiences [47], and do not therefore have this step.

Adaptation: In this step, the system attempts to determine if an adaptation should be performed, and decides upon the form of the adaptation.

Application of adaptation: Now that the system has determined that an adaptation needs to take place, the system performs the necessary transformations (adding, removing or colouring links or content, for example). In the case of serving an online, active user, the results of the user's request are delivered back to the user, transformed by the system's knowledge of the user.

Feedback: At this stage, the adaptation has been performed and it would be beneficial to measure the benefit (or lack of benefit) that it produces. This is not available

in many systems, because it is difficult to measure empirically, and feedback from users is scarce.

4.2 Common components

There are some common components which are found in most adaptive web system solutions.

Server: Almost all the systems involve capturing information from a web server, and the pages are eventually delivered by a web server. This component is responsible for recording user actions on the web site in addition to receiving and fulfilling web page requests. It is also involved in the identification (if any) of individual users.

User model: Most systems represent users via a user model. It represents both individual users and groups into which users are classified. It combines user preferences (stored in a user profile) with the stated goals or interests and the behaviours performed by that user, and uses this information to deduce the perceived current goals and interests of the user [56]. From this, it then will be used to try to predict what behaviour the user is likely to perform (or would perform, if the user had complete knowledge of all the possible choices available to them).

There are different types of user modeling system: individual versus canonical and static versus implicit acquisition [38]. For the most part, machine learning techniques are used to build the user model. Techniques used include linear models, TFIDF-based models, Markov models, neural networks, classification and clustering techniques, rule induction techniques and Bayesian theory-based techniques.

Domain model: While the users are represented by a user model, the website contents and structure are represented by a domain model. It includes not only the physical pages, construction rules (for a dynamic website, for example) and physical layout of the site, but may also include some semantic understanding of the layout of the website (whether explicitly stated by the web site designers or deduced using a machine learning technique). Often, the usage of the website is modeled directly over the domain model.

Adaptation model: In most systems, there is a third model which is usually present but not specifically articulated: this model represents the rules associated with the kinds of changes that can be made to the pages or the site. For the purposes of this discussion we label this model the *adaptation model*. Obviously, it will likely be closely associated with the domain model, but it also adds general decisions (such as style or presentation) that would be independent of the information domain presented in the domain model.

While these abstractions are useful for discussion, it must be recognized that in reality each is intimately related, and the choices for each impact on the necessary choices for the other. For example, the size of chunks of information that are present in the domain model limit the level of adaptability to which the user model might adjust, as well as limiting the kinds of changes the adaptation model might undertake.

4.3 Determining Users, Sessions and Transactions in the Server Log

In order to overcome the lack of information problems discussed earlier, it is necessary to pre-process the available data and infer additional information that is missing to make the record complete. This pre-processing is mainly done with the server logs, which are the primary source of activity information. In fact, information from other sources, such as the web pages themselves, can be considered as a source of additional information to be tied in with the server log data.

The data logs need to be analyzed in order to determine separate sessions which represent the interaction of a single user with a website over a particular period of time. Session discovery can be achieved by first partitioning the requests by the IP address; in this case, we simply assume each user has a unique IP address. Due to the usage of proxy servers and multiple users sharing the same IP address, an IP address is not a unique enough indicator of a session, so a further step must be taken. One technique is to use an inter-access delay threshold to split a sequence of requests from the same IP address [51].

Another approach to discovering sessions is described in [15]. In this approach, sessions are described as a set of transactions. In this case, there are two ways that transactions are defined, based on how a user treats web pages. Users treat web pages as either useful for navigation or interesting for the content they contain. Using this concept, we can define different transactions. The first way is to define a transaction as all of the navigation references up to and including each content reference for a given user; this is called the navigation-content transaction (also referred to as the auxiliary-content transaction in [16]). From these transactions we can easily find a popular path or common traversal paths in the web site. The other way to define a transaction is as all the content references for a given user; this is a content-only transaction. These transactions can be analyzed to give sets of pages which are frequently viewed together in a single session.

In order for sessions to be considered complete, they should show the complete path of page views from the beginning of the session to the end. As already mentioned, factors such as caching and backtracking may cause an incomplete path to be recorded in the log. Because of this, there must be some path completion algorithm used to fill in the missing information. The following techniques each attempt to do this:

Maximal forward reference (MFR): The MFR algorithm uses knowledge about the structure of a site to determine the longest sequence of page views in a session in which each page view may be reached via a direct link from the previous one. Once this is done, it may either decide that this is simply one transaction or a complete session on its own. In this case, missing pages in the log cause the transaction or session to be fragmented [14].

Shortest path (SP): The SP algorithm simply seeks to complete the path between two page views with the shortest path between them, if possible (and within limits). It follows an Occam's Razor assumption that a shorter path is always better and more accurate than a longer path [51].

Popular path (PP): The PP algorithm fills in gaps in the path between two page views with the complete path between these two pages which has most frequently been visited by users [51].

Mined path (MP): The MP algorithm completes paths by examining all the possible sub-paths (or path fragments) between the two pages and using the path with the best combination of high-popularity sub-paths [51].

4.4 Analyzing Web Page Content

In order to discover user interests, we need ways to compare the pages they have visited. A number of methods deriving from information theory have been used to compare web pages for similarity.

4.4.1 TFIDF

Term Frequency-Inverse Document Frequency (TFIDF) is one method of categorizing text documents which is commonly used [26, 17, 13]. In this method, important terms in the text are determined by calculating the number of times that term is found in documents within a category (term frequency), factoring in how rarely that term occurs over all documents (inverse document frequency).

Once these document categories are established, the process of categorizing a new document is as follows:

1. The existing categories and the new document needed to be classified are presented as vectors that consist of keywords with their frequencies. The keywords are distinct words or word parts which distinguish the existing categories or are prominent in the new document.

2. The inverse document frequency for each keyword is calculated; this is simply the reciprocal number of documents in which the keyword occurs at least once. The inverse document frequency of a keyword is high if it occurs in only one document, which makes it more significant for identifying the category of the document.

3. The weight value for each keyword is calculated by multiplying its frequency and its inverse document frequency.

4. The similarity between the new document and each existing category is calculated by computing the cosine value between their representative vectors.

5. The new document is placed into the category with the highest similarity.

4.4.2 Naïve Bayes Classifier

Naïve Bayes is another text categorization algorithm. Each existing category is represented as a vector of the probabilities of a given list of keywords appearing in documents in that category. To classify a document the probability of that document belonging to each category is calculated with the probability of each of its keywords in each class. Finally, the document is placed in the category with the highest probability. In this algorithm, all keywords are assumed to be independent of each other, which is to say that the appearance or absence of one keyword does not influence the likelihood of the appearance or absence of another [26].

4.4.3 PrTFIDF

Probabilistic TFIDF (PrTFIDF) is a classifier derived from TFIDF. PrTFIDF provides a way to distinguish a document and the representation of the document. First, this algorithm uses a function to map the document to its representation with a certain probability according to the function. The function is chosen by the users to assign relevance judgments to documents. Then, the classifier uses the representation to classify a new document by following similar steps to TFIDF [26].

4.4.4 WAKNN

Weight Adjusted k-Nearest Neighbour Classifier (WAKNN) [24] is another efficient and commonly used algorithm for text categorization. The main steps of WAKNN are:

1. Build characteristic vectors for the documents in each category and the new document. Each vector contains the frequencies of the same set of keywords.

2. Normalize the word frequency in each vector in order to account for the difference in document lengths.

3. In all categories, find the k-nearest neighbours of the test document by calculating the cosine similarity.

4. Sum up the similarity to the k neighbours according to their class labels.

5. Categorize the document to the class with the highest similarity sum.

4.5 Dealing with Concept Drift/Shift

Klinkenberg and Renz [29] use three metrics, *accuracy*, *precision*, and *recall*, as indicators of concept drift in text classification problem. They compared these performance metrics in eight different learning methods combined with four data management approaches (full memory, no-memory, fixed size window and dynamic size window). They also investigated abrupt interest change (concept shift) by using the same metrics. The comparison shows that these three performance measures are well suited as indicators for both concept drift and concept shift. Using a window management technique leads to significant performance improvements for these learning methods. Moreover, using the adaptive window management approach gains further performance improvements.

Billsus and Pazzani [8] establish a two-model approach. It consists of two separate models, a short-term interest model and a long-term interest model. The k-nearest neighbour algorithm and naive Bayesian classifier were used for modeling short-term and long-term interest models respectively. A multi-strategy learning approach that deals with both the user's long-term and short-term interests was developed. It attempted to use the short-term first. If short-term interests are not satisfied, a long-term interest model is used. This mechanism enables their system to adapt to a user's changing interests.

Lam and Mostafa [34] use a Bayesian approach in the information filtering systems to track user interest changes. In their work, a combined approach of reinforcement learning and Bayesian interest tracker is used to perform user profile initialization and maintenance. Simulation studies show that this concept drift modeling can effectively detect

user interest changes. Furthermore, it is able to track the changes of user interests online.

Koychev [32] considers that a user's experience does not represent simply one set of interests, but rather a set of user interest contexts which may occur more than once. When a user visits pages, these pages are examined to determine if the user is still in the same interest context, has fallen back to a previous interest context, or has started a new context. This can help to strengthen recommendations based on the retrieved context rather than the single page view.

4.6 Graph Theory Approaches

It is natural to think of a website (or the Web in general) in terms of a graph and use graph theory and clustering techniques to analyze it.

4.6.1 Structure-Based Graphs

It is possible to construct a model of the relationships between web pages based on a graph representation of the web pages. In this graph, the edges represent the associations between the pages, and the vertices each represent individual pages.

By analyzing this structure, it is possible to discover additional information. [28] describes the search for *authorities* and *hubs* done in this manner. Briefly, an authoritative page has been recognized as a good source of information by multiple referrers; a hub is a good referrer of multiple authoritative pages. By the application of this idea to the structure of a subset of pages returned from a query to a search engine, it has been demonstrated that one can determine good authorities and hubs in the graph. Authorities then make good recommendations for a user.

While this technique does discover good authorities and hubs, without some factoring in of the actual content of the pages the discovered authorities and hubs may have little to do with the original query or interest. This is due to the highly connected nature of the Web. For example, the top index page of a website may be flagged as a good hub because it points to the core pages of several different topics; each authority that this hub points to need not have the same topic, which means that this is really not a good hub. Another similar example occurs when considering the role of search engines, which would appear as good hubs because they point to a large number of pages.

4.6.2 Usage Graphs

A graph of the usage of a set of web pages can also be drawn, with the edges no longer representing explicit links between pages but rather a measure of how often these pages are viewed together in the same session. This is an application of the idea of association rule, through which "interesting" co-occurrences between sets of items are discovered. Again, the inclusion of a similarity measure of the content of the transactions can assist [40, 42, 43, 57].

Graphs of usage can be used as a way to represent individual users [12, 33]. By clustering these graphs and comparing an individual user to the cluster, recommendations can be drawn out of those pages which are part of the paths of the cluster but are not present in the user's own usage graph. An approximation of a user's interests (for example the page interest estimator (PIE) [12]) can improve the relevance of the results.

4.6.3 Association Rule Hypergraphs

An association rule is a concept which describes the relationship between a number of items (or a number of sets of items) with respect to their co-occurrence; that is, it describes the likelihood that a certain set of things will happen at the same time [2].

For the purposes of adaptive websites, association rules have a lot of applicability: they can be used to analyze the likelihood of the co-occurrences of two or more specific pages within particular visits; on a more abstract level, they can be used to analyze the co-occurrence of transactions within a session.

Association rules may be analyzed as hypergraphs, where each vertex represents a page view or transaction, and each edge connects to each page view or transaction that is associated together by a rule. Each edge carries a strength of association (also described as a weight of association) based on the support and confidence of that rule.

Finally, by clustering together vertices which are tightly related to each other, one discovers commonalities of usage which can be used to characterize individual users. By computing the most similar cluster to the user's own page views or transactions, recommendations can be given to likely pages that the user will visit [25, 41, 23].

5. IMPLEMENTATIONS

In this section, some of the more prominent or interesting implementations are discussed. This list is not meant to be complete or comprehensive, only illustrative.

5.1 WebWatcher

WebWatcher [27] is a software agent built into the website of the School of Computer Science of Carnegie Mellon University. It was in operation from August, 1995, to February, 1997 and worked as a tour guide for the Web. The agent has three main features:

1. It is familiar with the structure of the website, which includes the contents of each web page and links among the pages.

2. The agent can interact with users using natural language.

3. It learns by observing the actions of users. The more knowledge about users it has, the higher quality of recommendations it can provide.

Once the WebWatcher is activated on the Front-Door page of CMU School of Computer Science, WebWatcher commands are labeled on the top of each web page along with relative links listed in the middle of the page. The hyperlink URLs in the original page are rewritten to point to the WebWatcher server, with a tag to refer to the original destination. No matter whether the user selects recommended links are chosen or not, the WebWatcher accompanies the user to the next page and watches the user's actions to provide recommendations. Aside from this, WebWatcher can help users to search relative pages quickly.

WebWatcher implements the technologies of adaptation of content and navigation. It speeds up users' search and help user to avoid being lost in hyperspace. In addition, it attracts users as well.

5.2 SeAN

SeAN [3] is an adaptive system for the personalized access to news servers on the Web. SeAN provides personalized news items of topics which match a user's interests and an adaptive level of detail for news items.

In SeAN, news items are classified according to an *a priori* hierarchy of topics, and stored in a news database. Each news item is separated into several complex composite entities, associated with a few attributes that define their components. This structure enables the system to provide different levels of detail to match the knowledge and background of users.

Advertisements are stored in a advertisement database and each advertisement has two associated attributes: a target, which is the kind of user at which it is aimed, and a topic, which is the news section with which the advertisement is associated.

User profiles are stored in a user database. A user profile is divided into four areas: interests, expertise, cognitive characteristics, and life style. Each area corresponds to different conceptual characteristics of a user.

Stereotype profiles are used to initialize a user's profile. A stereotype profile consists of a set of user features and a probability that a user matching those features belongs to that stereotype. A user's responses to a registration form are compared to the stereotypes to compute the probability that the user belongs to each stereotype. Using this information, the system predicts the initial values of the rest of the user features using Bayes' Theorem. After initial creation, the user profile is dynamically updated according to the user's behaviour and user modeling update rules.

Content and presentation are modified to provide adaptive news in SeAN, including optional detail display and fragment variants. All pages are dynamically generated according to the user's profile, including the home page. Users with different interests and expertise will then reach different topics with different detail levels. In order to give a user some control over the decisions made by the system, the user can manually add or remove topics from their user profile, overriding the system's choices. These actions provide explicit feedback to the system, which can be used to correct the system's perceptions of the user.

5.3 AHA

The AHA [19] system is aimed at being a generic adaptive system which can be used for a wide variety of applications. It uses a user model which tracks the level of knowledge of particular concepts and uses that to decide if particular page fragments or links should be displayed, altered, or omitted. Each concept (which can also be described as an interest or preference) has a boolean value, which represents 'understood/not understood' (or 'interested/not interested', or 'preferred/not preferred'). A user's familiarity with concepts is drawn either from a test (or questionnaire) or by analyzing which pages they have viewed (pages may indicate that reading them, the user now understands a concept).

Pages are divided up into fragments, each wrapped with hidden, commented sections which are surrounded by 'if-then-else' clauses with the concept(s) that must be understood in order to view that section. Entire pages may also have dependencies, which indicate which concepts must be understood in order to view the page at all.

Links are also annotated with some indication of whether it is appropriate for the user to follow them, based on whether the page in an external link, or an internal link to a page with understood dependencies. The AHA server rewrites all URLs to be redirected to itself. In this way, it controls all interactions with the user.

The AHA system has intentionally been kept simple in order to broaden its applicability and create a standard platform for adaptive systems. It is limited by the need to explicitly label and annotate all of the pages over which the system is to have intelligence.

5.4 Other Implementations

There are a number of other implementations which bear mentioning. This list is not exhaustive.

- *GAS (Group Adaptive System)* [5]: A system that proposes to combine authority and hub search (([28]) with the idea of *social* or *collaborative navigation*, where individuals can provide navigational support for other individuals.
- *WEBMINER* [15, 16]: Research for this system focuses on different approaches to cleaning, completing, mining and analyzing web log data to provide high quality recommendations for users.
- *PageGather* [47]: This system generates potential index pages (subject to administrator approval) based on clustering a website's usage patterns.
- *Fab* [4]: This system uses a set of collaborating agents to provide recommendations for a user based on the user's interests derived from past page views.
- *Syskill & Webert* [7, 45]: This system categorizes pages retrieved from a search engine as to how interesting they would be to a user, based on the way that that user has explicitly rated pages in the past. Pages are rated as hot and cold by a user, and compared using a text comparison algorithm.

6. CONCLUSION AND SUMMARY

This paper has presented an overview of the goals, challenges, approaches and implementations that surround adaptive website research. This work is meant to provide an introduction to many of the most important difficulties, characteristics and solutions that have occurred to date, but is not intended to be an exhaustive overview. Readers are directed to [9, 11, 16, 31] for additional good overviews of the topic.

The title of this paper suggests the nature of the problem as both an art and a science. While considerable research has been performed into studying various aspects of the problem (how users behave in a web environment, the relationships between pages, and how to rate and rank suggestions to users, for example), there is still considerable art involved in producing effective adaptive web systems. This art ranges from the choices of particular parameters of clustering algorithms to how user interests may be measured or inferred, to measuring the effectiveness of a particular adaptation.

Nonetheless, substantial work has been done to explore the problem from three basic directions: understanding users, understanding websites, and understanding information. Most

approaches seem to examine the problem from one of these directions; some examine it from two of these directions; very few consider all three. We are in the midst of the early stages of the problem, where there is primarily *analysis* being performed, with the problem not yet sufficiently explored to allow broader *synthesis* to occur.

It is expected that for a substantial portion of time there will remain a large portion of the problem which can only be solved via the sound and steady application of considerable art, backed by the driving solidity of science.

7. ACKNOWLEDGMENTS

This research was funded in part by the Atlantic Canada Opportunity Agency (ACOA) through the Atlantic Innovation Fund (AIF).

8. REFERENCES

[1] Altavista's Babel Fish translator. `http://babelfish.altavista.com/`.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *Journal of the ACM SIGMOD*, pages 207–216, 1993.

[3] L. Ardissono, L. Console, and I. Torre. An adaptive system for the personalized access to news. *AI Communications*, 14(3):129–147, 2001.

[4] M. Balabanovic. An adaptive web page recommendation service. In *Proceedings of the 1st International Conference on Autonomous Agents*, pages 378–385, Marina del Rey, CA, USA, 1997.

[5] M. Barra, P. Maglio, A. Negro, and V. Scarano. GAS: Group Adaptive System. In P. De Bra, P. Brusilovsky, and R. Conejo, editors, *Adaptive Hypermedia and Adaptive Web-Based Systems, Second International Conference, AH2002*, volume 2347 of *Lecture Notes in Computer Science*, pages 233–241, Málaga, Spain, May 29-31 2002. Springer-Verlag.

[6] P. Batista and M. J. Silva. Mining web access logs of an on-line newspaper. In *Proceedings of the Workshop on Recommendation and Personalization in eCommerce of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*, Málaga, Spain, May 29-31 2002.

[7] D. Billsus and M. Pazzani. Revising user profiles: The search for interesting web sites. In *Proceedings of the Third International Workshop on Multistrategy Learning*. AAAI Press, 1996.

[8] D. Billsus and M. J. Pazzani. A hybrid user model for news story classifcation. In J. Kay, editor, *User Modeling: Proceedings of the Seventh International Conference, UM'99*, pages 99–108. Springer, 1999.

[9] P. Brusilovsky. Methods and techniques of adaptive hypermedia. *User Modelling and User-Adapted Interaction*, 6(2-3):87–129, July 1996.

[10] P. Brusilovsky. Efficient techniques for adaptive hypermedia. *Intelligent Hypertext: Advanced Techniques for the World Wide Web*, 1326:12–30, 1997.

[11] P. Brusilovsky. Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 11:87–110, 2001.

[12] P. Chan. Constructing web user profiles: A non-invasive learning approach. In *KDD-99 Workshop on Web Usage Analysis and User Profiling*, pages 7–12, San Diego, CA, USA, August 15-18 1999.

[13] L. Chen and K. Sycara. Webmate: A personal agent for browsing and searching. In *Proceedings of the 2nd International Conference on Autonomous Agents and Multi Agent Systems(Agents'98)*, pages 132–139, May 1998.

[14] M. S. Chen, J. S. Park, and P. S. Yu. Data mining for path traversal patterns in a web environment. In *Sixteenth International Conference on Distributed Computing Systems*, pages 385–392, 1996.

[15] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the World Wide Web. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, November 1997.

[16] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining World Wide Web browsing patterns. *Knowledge and Information Systems*, 1(1):5–32, 1999.

[17] B. D. Davison. Predicting web actions from html content. In *Proceedings of the ACM Conference on Hypertext*, pages 159–168, 2002.

[18] P. De Bra. Design issues in adaptive web-site development. In *Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW*, Toronto, Canada, May 1999.

[19] P. De Bra and L. Calvi. AHA: a generic adaptive hypermedia system. In *Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia, HYPERTEXT'98*, Pittsburgh, USA, June 20–24 1998.

[20] D. Dhyani, W. K. Ng, and S. S. Bhowmick. A survey of web metrics. *ACM Computing Surveys*, 34(4):469–503, December 2002.

[21] J. Eklund and R. Zeiliger. Navigating the web: Possibilities and practicalities for adaptive navigational support. In *Proceedings of the Second Australian World Wide Web Conference, AusWeb96*, 1996.

[22] J. Fink, A. Kobsa, and A. Nill. *User-oriented adaptivity and adaptability in the AVANTI project*, pages 135–143. Microsoft Usability Group, Redmond, WA, USA, 1996.

[23] A. Gyenesei. A fuzzy approach for mining quantitative association rules. Technical Report 336, Turku Centre for Computer Science, Turku, Finland, May 2000.

[24] E.-H. Han, G. Karypis, and V. Kumar. Text categorization using weight adjusted k-nearest neighbor classifier. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1999.

[25] E.-H. S. Han, G. Karypis, V. Kumar, and B. Mobasher. Clustering based on association rule hypergraphs. In *Proceedings of the SIGMOD'97 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'97)*, May 1997.

[26] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 143–151, 1996.

[27] T. Joachims, D. Freitag, and T. Mitchell. WebWatcher: A tour guide for the World Wide Web.

In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 770–775. Morgan Kaufmann, August 1997.

[28] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, September 1999.

[29] R. Klinkenberg and I. Renz. Adaptive information filtering: Learning in the presence of concept drifts. In M. Sahami, M. Craven, T. Joachims, and A. McCallum, editors, *Workshop Notes of the ICML-98 Workshop on Learning for Text Categorization*, pages 33–50, Menlo Park, CA, USA, 1998. AAAI Press.

[30] A. Kobsa. Personalized hypermedia and international privacy. *Communications of the ACM*, 45(5):64–67, 2002.

[31] A. Kobsa, J. Koenemann, and W. Pohl. Personalized hypermedia presentation techniques for improving online customer relationships. *The Knowledge Engineering Review*, 16(2):111–155, 2001.

[32] I. Koychev. Learning about users in the presence of hidden context. In R. Schfer, M. E. Mller, and S. A. Macskassy, editors, *Proceedings of the UM2001 Workshop on Machine Learning for User Modeling*, pages 49–58, Sonthofen, Germany, July 13-17 2001.

[33] D. Kukulenz. Prediction of navigation profiles in a distributed internet environment through learning of graph distributions. In P. De Bra, P. Brusilovsky, and R. Conejo, editors, *Adaptive Hypermedia and Adaptive Web-Based Systems, Second International Conference, AH2002*, volume 2347 of *Lecture Notes in Computer Science*, pages 233–241, Málaga, Spain, May 29-31 2002. Springer-Verlag.

[34] W. Lam and J. Mostafa. Modeling user interest shift using a bayesian approach. *Journal of the American Society for Information Science and Technology*, 52(5):416–429, March 2001.

[35] S. Lawrence and C. L. Giles. Searching the World Wide Web. *Science*, 280(5360):98–100, 1998.

[36] H. Lieberman. Letizia: An agent that assists web browsing. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 924–929, San Mateo, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[37] N. Mathe and J. Chen. A user-centered approach to adaptive hypertext based on an information relevance model. In *Fourth International Conference on User Modelling*, pages 107–114, 1994.

[38] M. F. McTear. User modeling for adaptive computer systems: a survey of recent developments. *Artificial Intelligence Review 7*, pages 157–184, 1993.

[39] D. Mladenic. Personal WebWatcher: design and implementation. Technical report, Department of Intelligent Systems, J. Stefan Institute, Slovenia, 1996.

[40] B. Mobasher, R. Cooley, and J. Sravastava. Creating adaptive web sites through usage-based clustering of URLs. In *IEEE Knowledge and Data Engineering Workshop (KDEX 99)*, November 1999.

[41] B. Mobasher, H. Dai, T. Lou, and M. Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6:61–82, 2002.

[42] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Y. Sun, and J. Wiltshire. Discovery of aggregate usage profiles for web personalization. In *Proceedings of the WebKDD Workshop*, 2000.

[43] B. Mobasher, H. Dai, T. Luo, Y. Sun, and J. Zhu. Integrating web usage and content mining for more effective personalization. In *Proceedings of the International Conference on E-Commerce and Web Technologies*, pages 165–176, Greenwich, UK, September 2000.

[44] J. Nielsen. User interface directions for the web. *Communications of the ACM*, 42(1):65–72, 1999.

[45] M. J. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27(3):313–331, 1997.

[46] M. Perkowitz and O. Etzioni. Adaptive web sites: An ai challenge. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pages 65–72, May 1997.

[47] M. Perkowitz and O. Etzioni. Adaptive web sites: Automatically synthesizing web pages. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Madison, WI, USA, 1998.

[48] I. Schwab, A. Kobsa, and I. Koychev. Learning about users from observation. In *Adaptive User Interfaces: Papers from the 2000 AAAI Spring Symposium*, Menlo Park, CA, USA, 2000. AAAI Press.

[49] I. Schwab, W. Pohl, and I. Koychev. Learning to recommend form positive evidence. In *Proceedings of the 2000 International Conference on Intelligent User Interfaces*, pages 241–248, New Orleans, LA, USA, 2000.

[50] R. Srikant and Y. Yang. Mining web logs to improve website organization. In *Tenth International World Wide Web Conference*, Hong Kong, May 1-5 2001.

[51] F. Toolan and N. Kushmerick. Mining web logs for personalized site maps. In *Workshop on Mining for Enhanced Web Search, International Conference on Web Information Systems Engineering*, 2002.

[52] M. Virvou, V. Tsiriga, and M. Moundridou. Adaptive navigation support in a web-based software engineering course. In *Proceedings of the 2nd International Conference on Technology in Teaching and Learning in Higher Education*, pages 333–338, 2001.

[53] G. Weber and M. Specht. User modeling and adaptive navigation support in WWW-based tutoring systems. In A. Jameson, C. Paris, and C. Tasso, editors, *Proceedings of the Sixth International Conference of User Modeling, UM'97*, pages 289–300. Springer-Verlag, 1997.

[54] P. Wolfgang, A. Kobsa, and O. Kutter. User model acquisition heuristics based on dialogue acts. In *International Workshop on the Design of Cooperative Systems*, pages 225–226, 1995.

[55] H. Wu, G.-J. Houben, and P. De Bra. Aham: A reference model to support adaptive hypermedia authoring. In *Proceedings of InfWeb 98*, 1999.

[56] H. Wu, G.-J. Houben, and P. De Bra. Authoring support for adaptive hypermedia applications. In *ED-MEDIA conference*, pages 364–369. AACE, 1999.

[57] Y.-H. Wu, Y.-C. Chen, and A. L. P. Chen. Enabling

personalized recommendation on the web based on user interests and behaviors. In *Proceedings of the International Workshop on Research Issues in Data Engineering - Distributed Object Management -RIDE-DOM*, pages 17–24, 2001.

[58] J. Zaslow. If TiVo thinks you are gay, here's how to set it straight. *The Washington Post Online*, November 2002.

APPENDIX

A. DEFINITION OF TERMS

adaptable A web system is merely *adaptable* if the way it performs or behaves changes based on explicit information, such as a user profile. In addition, this profile will not vary over time unless the user explicitly changes it [18].

adaptive A web system is considered *adaptive* if it changes based on implicitly discovered information, such as an analysis of the way it is used. Usually, the instrument of change is somewhat automated, but in some cases, human intervention is used. Contrast with *adaptable* and *dynamic*.

aggregate usage profile One of the most prominent techniques in all machine learning appears to be the clustering of information into groups of high similarity. In adaptive web systems, it is useful to cluster the way that the site is used into common groups which share patterns. These groups are called *aggregate usage profiles*.

atom The smallest possible self-contained section of a web page. Examples include: a single new article; an advertisement; an image; a link; a navigation bar. In a more general sense, an atom consists of information which is all of a single, consistent concept [18].

closed-corpus Literally, "closed body" of web pages. This is describes a set of web pages within a particular boundary, such as a single website, and is used to contrast against systems which operate over a known dominion of known pages rather than systems which operate on the "open" collection of web pages that is the Web in general. Some techniques [28] create a closed-corpus set of pages by using another selection tool (such as a search engine) to pull out a subset of pages.

dynamic A web system is *dynamic* if it delivers pages which are composed at request-time of a number of smaller components or *atoms*. The rules to make this composition may also be partially determined at page request time based a user profile.

page view The most basic form of website usage is the viewing of an individual "page". *Page* in this context refers not necessarily to the viewing of a particular, static HTML document, but really refers to the viewing of a particular, discrete set of atoms [18].

session A item of measurement for website usage which describes collections of transactions for a particular user. A session is usually defined in terms of a window of time in which that particular user was active and interested in a particular topic.

static A website is *static* if it presents the same information every time to every user. Most website do change over time, but the view that each user receives is not modified or informed by who that user is.

transaction An item of measurement for website usage which focuses on "interesting" collections of events for a particular user. Usually, a transaction must be inferred from the usage data rather than simply retrieved from it. The particular definition of a transaction varies from the viewing of a page to the selection of an item (as in an e-commerce system), and is dependant on the particular adaptive solution one is employing.

user interest A description of the topics or goals for which the user is looking. In the case of a search engine, for example, the user's goal is to find pages which contain topics related to their search query. One assumption that is made about user interests is that they can be inferred from a combination of user information, browsing patterns and website data.

user profile A collection of information about a user, combining demographic information (name, age, location, for example), usage information (pages visited, frequency of visit, for example), and interests or goals (either explicitly stated by the user or implicitly derived by the system).