A Framework for Self Adaptive Websites: Tactical versus Strategic Changes

Filip.Coenen, Gilbert.Swinnen, Koen.Vanhoof, Geert.Wets@luc.ac.be

Limburg University Centre, Faculty of Applied Economic Sciences, B-3590 Diepenbeek, Belgium

Abstract

In the contemporary information society where the Internet performs a leading role, and the competition is just one mouse-click away, it is very important to communicate your message to the Website visitors in an optimal way. When creating a Website, the designer meticulously creates the look and feel of the site, designs the structure of the information, and determines the kinds of interactions that must be available. However, in order to serve the Website visitor better, a lot of work has already been devoted to the development of self adaptive Websites, where the original site structure is of secondary importance. In this paper we introduce a framework for the implementation of adaptive Websites and distinguish between tactical and strategic adaptations that can be made to a Website. Subsequently we introduce a technique that covers the problem of the tactical changes being made by adaptive Websites.

Keywords: Self-adaptive Websites, Tactical Adaptations, Strategic Adaptations.

1. Introduction and problem description

A lot of contemporary research is being carried out in the field of analyzing and understanding the user behavior on a Web site. In this context, a variety of data mining techniques have been proposed, including the discovery of association rules and sequential patterns, and clustering (Han et al, 1996) as will be explained in the related work section.
Once these techniques have revealed certain patterns, these can be used in the construction of Web sites that automatically improve their organization and presentation by learning from the Website user's online behavior. These Web sites are referred to as self-adaptive Web sites (Etzioni and Perkowitz, 1997), where a partial session for the current user can be assigned to a matching pattern. When a surfer is visiting a self-adaptive Website, other URLs from such a matching pattern can be provided to him as recommendations for further surfing on the Web site. The problem evolving from this approach is twofold. First, it may be hard for the visitor to keep overview of the offered recommendations. Second, due to the nature of a self adaptive Website, it may be hard for the visitor to not get lost in hyperspace, since nothing guarantees that he can easily find his way back to the desired information when the site is adapted in between.
In this paper we introduce a framework for the implementation of adaptive Web sites and subsequently, we introduce a technique that covers the problem of the tactical changes being made by adaptive Web sites.

2. Related Work

As is mentioned in the introduction, a variety of Data mining techniques have been proposed for the analysis of user behavior on a Website, the most popular of which are the discovery of association rules and sequential patterns, and clustering (Han et al, 1996). Association rule discovery techniques are generally applied to databases of transactions where each transaction consists of a set of items (Agrawal and Srikant, 1994). In the context of Web mining, the problem amounts to discovering the associations among the URLs that appear in Web transactions or sessions. In this context, a Web transaction or session is defined as a series of pages requested by a same user during a single visit to a Web site. The discovery of these association rules is based on the determination of frequent item sets, i.e. sets of which the support, or the number of occurrences, exceeds a given threshold.
Agrawal and Srikant (1995) introduced some algorithms for the discovery of sequential patterns in large databases. An example of such a pattern or sequence might be that typically event A happens, followed by event B, and then event C. Again, when the support of the sequence is larger than a certain threshold, the sequence is called to be frequent. An important feature of such a frequent sequence is that the events in a frequent sequence need not to happen consecutively, i.e. a sequence where other events happen in between, e.g. event A – event D – event B – event C, also supports the sequential pattern A – B – C. It is the application of this technique to Web server access logs that allows Web-based organizations to predict user visit patterns.
Equally, clustering algorithms can be used to discover hidden common patterns among data items.

Specifically, the term clustering refers to the data mining activity that allows one to group together data items that have similar characteristics (Kaufman and Rousseeuw, 1990). This technique can also be applied in order to compute clusters of URL references that are similar (transaction clusters) or for the direct computation of clusters of URL references based on how often they occur together across user transactions (usage clusters).

Equally, a lot of research has been done as far as the human information processing is concerned. Jacob Nielsen's (1997) research on the reading behavior Web users resulted in the conclusion that users don't read, but instead scan the Web page. Therefore, Web pages have to employ scannable text, including keywords, sub-headings and bulleted lists. As a reason for this behavior, Nielsen quotes four plausible reasons:

- Reading from computer screens is tiring
- The Web is a user-driven medium, so people don't experience any productivity while not clicking
- Each page has to compete with hundreds of millions of other pages for the user's attention, so users don't know whether this page is the one they are looking for
- Due to the nature of today's hectic life, people simply don't have time to work hard for their information.

In his further research on user behavior on the Web, Nielsen comes to the following general conclusion as far as user characteristics are concerned: users look straight at the content area and ignore the navigation areas when they scan a new page. This means that users are extremely goal-driven, and only look for the thing they have in mind, thereby ignoring other aspects of design, advertising and other elements. As far as the actual navigation is concerned, Nielsen states that it is crucial to provide the users with a site map, and to let them know where they are and where they can go next. According to the information found on the site of CNET, it is indeed critical that users always know where they are in a Web site and that they are comfortable getting around. The goal of the site administrator therefore is to make it easy for users to get from page to page, and to make it immediately obvious for them how to get back home. Consequently, visitors should have cues that help them remember where they came from, and they should find clear suggestions on where to go next.

In 1956 already, Miller discussed 'The Magical Number Seven, Plus or Minus Two' in the context of the limits on the human capacity for processing information. In his experiments, the observer is considered to be a communication channel. The experimental problem is to increase the amount of input information, and to measure the amount of transmitted information. If the observer's absolute judgments are quite accurate, then nearly all of the input information will be transmitted and will be recoverable from his answers. If he makes errors, then the transmitted information may be considerably less than the input. He expected that, if the amount of input information were increased, the observer would begin to make more and more errors. When the human observer is a reasonable kind of communication system, the amount of transmitted information was expected to increase at first, and eventually level off at some asymptotic value. The conclusions from Miller's research are threefold. First, he claims there is a finite span of immediate memory of about seven items in length. Second, he concludes from his experiments that there is a span of absolute judgment that can distinguish about seven categories. His third conclusion concerns the fact that there is a span of attention that encompasses about six objects at a glance. A more specific conclusion as far as linguistic experiments are concerned, is supported by the results of Hayes' experiments (1954). Hayes concludes that most people can remember between 5 and 9 unrelated chunks of information, where a chunk can either be a letter, a word or a group of words. These conclusions are still supported nowadays, as e.g. can be quoted from William Maki (1998).

3. Proposed Approach

When creating a Web site, the designer meticulously creates the look and feel of the site, designs the structure of the information, and determines the kinds of interactions that must be available. Considering this, we believe that there should be made a severe distinction between strategic (long-term) changes on the one hand, and tactic (short-term) adaptations on the other hand when a self-adaptive Web site is implemented, since it is desired to avoid damaging the structure of the site when automatic changes are invoked. The danger exists that the surfer gets confused about his current location on the site when adaptations affecting the site structure are carried out in real time. The original Web site structure is not taken into account when the self-adaptive Web site issues recommendations based upon frequent item sets, frequent sequences or clusters due to the nature of these analysis results. This way, links would be proposed that belong to the same item sets, sequences or clusters, without taking into account if these links exist in the original site structure. Consequently, the beliefs of the site designer would be violated.

In this context, we define strategic adaptations as those adaptations that go against or conflict with the original beliefs of the site, and consequently have an important influence on the original site-structure.
Tactical changes are those changes that adapt the site according to its use, but leave the site structure unaffected. A typical example of a tactical adaptation would be the automatic creation of a customized recommendation list. When the theories discussed in the related work section are projected to the concept of Website design, this implicates that the optimal length of a link list is determined to be about seven, and ideally does not exceed a length of five. When this threshold of five is exceeded, the chances for the surfer to loose overview of the proposed information grow exponentially.

3.1 Framework for self-adaptive Web sites

3.1.1 Strategic adaptations

As mentioned before, the danger exists that the surfer gets confused about his exact location on the site when adaptations affecting the site structure are carried out in real time. The techniques that are discussed in the related work section (association rules, sequence rules and clustering) are the most commonly implemented techniques for the realization of an adaptive Web site (Han et al, 1996). However, due to the nature of the results of these methods, the original Web site structure is not taken into account. This means that when the self-adaptive Website issues recommendations based upon frequent item sets, frequent sequences or clusters, it proposes other links that belong to these item sets, sequences or clusters, not taking into account if these links exist in the original site structure. In this way, the beliefs of the site designer are violated. For example, a wholesaler goes online and has his product information organized in a categorical way by nature of the product, since this is his belief on his product assortment. Among other categories, he also offers products from the section animals and the section furniture. When a lot of visitors are interested in some goldfish that are listed in the section animals, and in some type of aquarium from the section furniture as well, these two items appear together in the frequent sets, frequent sequences or clusters. Consequently, when a new visitor hits the link goldfish, the link leading to the aquarium is suggested to him. Possible problems that arise when this approach is implemented are the following. The visitor is taken from one section to another, with the only link between them being a conceptual one. This implies that the theory concerning Web site navigation that is covered in the introduction is violated. Further, when the visitor wants to come back later to visit the aquarium page again, it is hard for him to get there directly, since the only way he knows to reach the considered page is via the goldfish page. Further, it is perfectly possible that other recommendations did come forward on the goldfish page, what makes that the aquarium link does not exist anymore on the considered page.
Therefore we believe that these strategic adaptations based upon the discovery of frequent item sets, frequent sequences and clusters should be made offline, thereby revising the overall original site structure. In the case of the wholesaler example this means that when the link concerning the goldfish and the link concerning the aquarium appear to make up an important cluster, they should be brought together under a new category 'fish friends' that must be accessible from the homepage. Since this category is a frequently visited section, it can be considered to expand the available information concerning fish and relevant accessories, in order to please this important group of visitors even more. In other words, the original beliefs of the Website administrator should be reviewed, thereby not forgetting the restricted link number of maximal seven, as stated in section 3.
As far as the tactical changes are concerned, we propose in the next section an approach for making recommendations to Web site visitors. This approach does not damage the original site structure, and in this way leaves unaffected the beliefs of the site administrator.

3.1.2 Tactical adaptations

Based upon the theories explained in the introduction, we believe that the only adaptations that should be triggered in real time by adaptive Web sites are tactical adaptations, i.e. those adaptations that do not affect the structure of the site. This means that the existing link structure must always be taken into account when recommendations are made to the visitor. A similar approach as the one we introduce next has been proposed by Schechter, Krishnan and Smith (1998). However, they showed how to predict request behavior using path profiles in order to let the server prefetch the information that the visitor is most likely to address next. In this way, the end user will witness significantly lower latencies for their requests. Equally, Tauscher (1997) proposes several approaches for the creation of so-called history lists, however the application is limited to the improvement of the browser 'back button'.
The approach we propose is to use the algorithm that is explained further in order to suggest on each page the most frequently visited links that previous

customers followed when they reached the same page.

Algorithm:
Initiate tree by root creation and setting support threshold
For each session do
Begin
    For each URL in session do
    Begin
        If URL exists in branch then go to node labeled URL and take next URL of session
        Else if (occurrence >= threshold) then add URL and calculate support
        Else create new branch and calculate overall support
    End.
End.

The purpose of the algorithm is to construct a tree containing all the different frequent paths that were calculated support. The first step in the algorithm is the initialization step and consists of the root creation of the tree. The support threshold is determined to be 2, so the root support must amount to 2 in order to allow the tree to grow. In the first iteration, the first URL of the session is considered. Since the support of the tree equals the threshold, a child node labeled A is created. The overall support of URL A amounts to 2, since this URL appears twice in the session. This support equals the threshold, so in iteration 2 the tree is allowed to grow with the next URL B. Though URL B appears twice in the session, the support is set to 1 in this node, since the sequence A-B appears only once in the session, i.e. B follows only once after A. In the third iteration the considered URL is again named A. Since this URL already appears in the branch, the node labeled A is considered as the current node. As mentioned before, the support of this node equals the threshold, so in iteration 4 a child node labeled C with a support that equals 1 can be created. Again, in iteration 5 the next URL to be evaluated is labeled B.

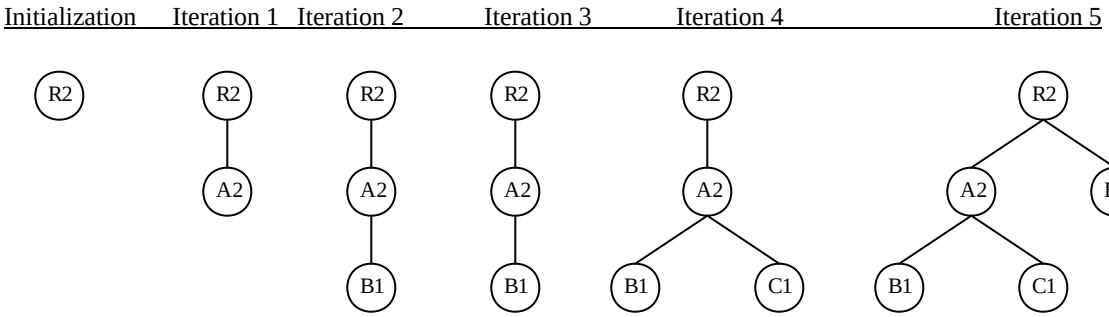

Figure 1: Example

followed by the visitors. In order to further explain this algorithm, we will first explain the conditions that we believe a frequent path should meet before it is inserted in the tree.
The tree is initiated by creating a root node with a certain support threshold. When recording a path in the tree, the first URL in a path will be stored as a child of the root node of the tree. The second URL in the path will be stored in a node that is a child of the first URL's node. This may continue until the complete path is stored in the tree. However, in order to reduce the number of paths in the tree, and to keep overview and avoid exceeding resource constraints, in every node the support of the path is checked against the initially given threshold. For example consider the case where the current session consists of the URL-sequence A-B-A-C-B, the threshold is set to be 2, the URL B is the current node in the tree and URL C is being evaluated. The drawings illustrate the nodes with the corresponding URL label and the However, B does not appear in the current branch, and since the support of the node labeled C is less than the threshold, a total new branch has to be initiated. The support of the new node labeled B is set to 2, i.e. the overall support of URL B in the session. This procedure has to be followed for all the available user sessions in order to build the complete tree.
When the final tree of figure 1is supposed to be the complete tree after the running of the algorithm on all the user sessions, it can be interpreted as follows. From the two times that page A is visited, one visit lead on to URL B and one visit lead on to URL C. After the construction of the tree, and taking into account the theories explained before, recommendations to the Website visitors can be made in an obvious way by proposing on each Website page its five best children according to the tree. In this way, the original site structure will stay unaffected. This implies that the Website structure

can be communicated to the visitor in a uniform way, in accordance with the theories on human information processing that are explained in the related work section. The tree also allows the Website constructor to evaluate the links on a Website page, and to conclude which links are the best ones and which should be reconsidered, something the existing sequence rule techniques are not capable of, as will be explained next. In this way, the concept of a self-adaptive Website can be applied while the risk for the visitor to get lost in hyperspace is minimized. This can easily be explained by making clear the difference between the proposed approach and the more traditional sequence rules. Suppose two visitor sessions are logged, being A-B-C-D and A-F-C-D. Based upon this example, the most frequent sequence according to the traditional sequence rules is A-C-D with a support amounting 2, which illustrates that the existing site structure is not taken into account. When the introduced algorithm is run over the sessions, the branches of the corresponding tree would be A(2)-B(1), C(2)-D(2), A(2)-F(1). Although it is important to consider the method of session determination (Cooley, 1999), the fact that the session is split into multiple branches has no impact on the purposes of our analysis, i.e. the making of recommendations. If recommendations were made, based on the sequence rules, the visitor arriving at A would be proposed to go to C, although this link did not exist in the original site structure. The visitors had to follow either B or F in order to surf from A to C as can be derived from the sessions. According to the proposed technique, a visitor arriving at A would be suggested to click on to either B or F, although he probably is more interested in C. The proposal in this case would be to use the sequence rules for enhancing the Website structure by means of a strategic change as is mentioned before, in this case by providing a link directly from A to C. The future recommendations that will be made online would then be based on the branch A(2)-C(2)-D(2). For reasons of completeness, a comparison between the Web usage miners WUM and MiDAS is mentioned here. With both miners, the possibility exists to take into account the existing site structure. However, in MiDAS, this step is implemented as an ex post approach, i.e. after the termination of the sequence discovery (Büchner, 1999). In WUM, a mining query language (MINT) is provided for the discovery of sequences that satisfy a desired template instead of a general sequence discovery approach (Spiliopoulou, 1999).

3.2 Implementation

Unfortunately, we have not been able so far to implement the algorithm in a real life website. The algorithm has only been implemented in a tool that is used the offline evaluation of existing Website links. The output of the algorithm was designed in such a way that it can be read by a 3D hyperbolic module called Hypview (Tamara Munzner), as is illustrated in figure 2.

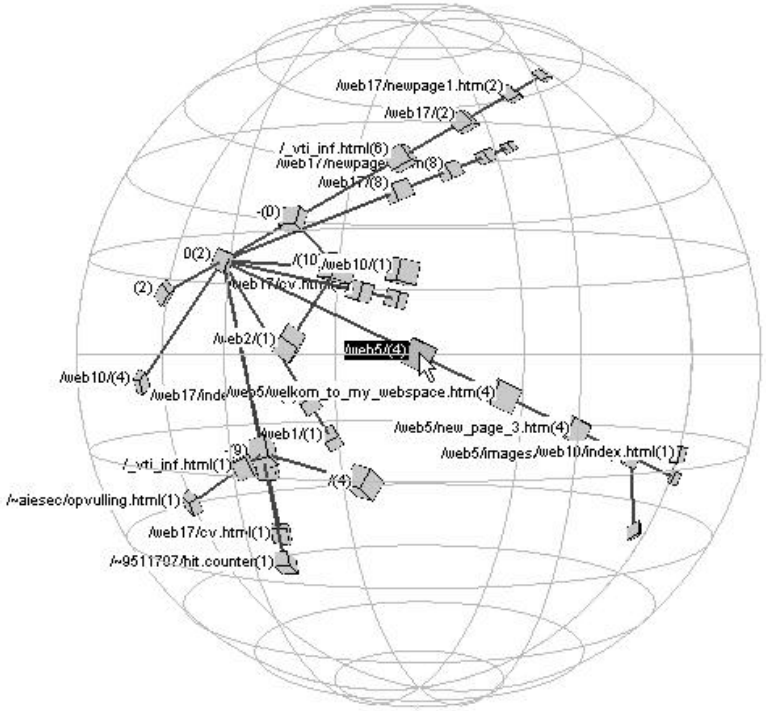

Figure 2: The algorithm output visualized by Hypview

Based on this application, it is very easy to evaluate existing links. When the node that is indicated by the cursor in figure 2 is taken for example, it can be seen that all four visitors who arrived at the URL '/web5/' clicked on to the URL '/web5/welkom_to_my_webspace'. Instead of enabling a visual inspection of the Hypview tree, our eventual purpose is to implement the outcome of the algorithm in the corresponding Website, and build a recommendation structure based upon this output.
Equally, in the near future our approach is to be compared to the functionalities of the footprints model proposed by Alan Wexelblat. Equally, this model makes recommendations for website visitors based on paths frequently visited by previous visitors.

4. Conclusions

Self-adaptive Websites are Websites that automatically improve their organization and presentation by learning from the Website user's online behavior based upon certain pattern revealing techniques such as the discovery of association rules and sequential patterns, and clustering. However, due to the nature of these techniques, the original site structure is not taken into account. Given this, we introduce in this paper a framework for self-adaptive Websites, based on our believe that there should be made a severe distinction between strategic and tactical changes that are invoked by the Website, in order to avoid the visitor from getting lost in hyperspace. The conclusions emerging from the techniques mentioned before should lead to offline adaptations, allowing changes to the Website structure. As far as the tactical adaptations are concerned, we introduced in this paper an algorithm for making online recommendations, thereby leaving unaffected the existing site structure and taking into account the existing theories on human information processing.

5. References

Agrawal, R., Srikant, R., Fast algorithms for mining association rules, in Proceedings of the 20th VLDB conference, pages 487-499, Santiago, Chili, 1994.

Agrawal, R., Srikant, R., Mining sequential patterns, in Proceedings of the 20th International Conference on Data Engineering, Taipei, Taiwan, 1995.

Agrawal, R., Srikant, R., Mining sequential patterns: generalizations and performance improvements, IBM Research Division.

Büchner, A.G., Baumgarten, M., Anand, S.S., Mulvenna, M.D., Hughes, J.G., Navigation Pattern Discovery from Internet Data, in Proceedings of the Workshop on Web Usage Analysis and User Profiling, WEBKDD '99, pp. 25-30, San Diego, California, 1999.

CNET,
http://www.builder.com/Graphics/CTips2/ss12.html

Cooley, R., Mobasher, B, Srivastava, J., Data Preparation for Mining World Wide Web Browsing Patterns, in Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX-97), November 1997.

Dayal, U., Garcia-Molina, H., Jacobsen, M., Yan, T. W., From User Access Patterns to Dynamic Hypertext Linking, in Computer Networks and ISDN Systems, volume 28, issues 7-11, pp. 1007.

Etzioni, O., Perkowitz, M., Adaptive Web Sites: an AI Challenge, In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, 1997.

Etzioni, O., Perkowitz, M., Adaptive Web Sites: Automatically Synthesizing Web Pages, AAI, 1998.

Han, E.-H., Jain, N., Mobasher, B., Srivastava, J., Web Mining: Pattern Discovery from World Wide Web Transactions, Department of Computer Science, University of Minnesota, Minneapolis, USA, Technical Report 96-050, 1996.

Hayes, J. R. M. Memory span for several vocabularies as a function of vocabulary size. In *Quarterly Progress Report*, Cambridge, Mass.: Acoustics Laboratory, Massachusetts Institute of Technology. Jan - June 1952.

Kaufman, L., Rousseeuw, P.J., Finding groups in data: an introduction to cluster analysis, John Wiley and Sons, 1990.

Maki, W,
http://ttacs7.ttu.edu:8902/psy1300/outlines/faq/faq7-2.html, 1998.

Miller, G.A., The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information, *Psychological Review*, *63*, 81-97, 1956.

Munzner, Tamara, Hypview, http://graphics.stanford.edu/~munzner/.

Nielsen, Jacob, How Users Read on the Web, http://www.useit.com/alertbox/9710a.html,1997.

Schechter, S., Krishnan, M., Smith, M.D., Using path profiles to predict HTTP requests, in Computer Networks and ISDN systems, vol. 30, issue 1-7, pp. 457-467, 1998.

Spiliopoulou, M., Pohle, C., Faulstich, L., Improving the Effectiveness of a Web Site with Web Usage Mining, in Proceedings of the Workshop on Web Usage Analysis and User Profiling, WEBKDD '99, pp. 51-56, San Diego, California, 1999.

Tauscher, L., Greenberg, S., Revisitation Patterns in World Wide Web Navigation, in the Proceedings of the International Conference CHI'97, Atlanta, Georgia, Mar. 1997.

Wexelblat, A., An Environment for Aiding Information-Browsing Tasks, in the Proceedings of AAAI Spring Symposium on Acquisition, Learning and Demonstration: Automating Tasks for Users, Birmingham, UK, 1996.