Using Adaptive Techniques to Validate and Correct an Audience Driven Design of Web Sites

Sven Casteleyn[1], Irene Garrigós[2], and Olga De Troyer[1]

[1] Vrije Universiteit Brussel, Department of Computer Science, WISE, Pleinlaan 2, 1050 Brussel, Belgium
{Sven.Casteleyn, Olga.DeTroyer}@vub.ac.be
[2] Universidad de Alicante, IWAD, Campus de San Vicente del Raspeig, Apartado 99 03080 Alicante, Spain
{igarrigos}@dlsi.ua.es

Abstract. An audience driven philosophy for web site design takes the different target audiences as an explicit starting point, and organizes the basic navigation structure accordingly. However, for the designer it is not always easy, or sometimes even impossible, to assess the different requirements of the different target audiences correctly. In this paper, we show how to correct for such possible flaws using adaptive behavior. A mechanism for detecting both missing and superfluous information in a certain user's navigation path, by observing user's browsing behavior, is provided. The designer specifies possible adaptive changes based upon this detection at design time, using a language (Adaptation Specification Language) designed specifically to express changes in the navigation structure of a website.

1 Introduction

One approach described in the literature to improve usability [10] of web sites is the Audience Driven design philosophy [5][6]. For a web site design, it takes as a starting point the identification of the different target audiences, and arranges them in a hierarchy according to their requirements. From this hierarchy, the main structure of the web site is derived. Concretely, for the visitors this results in links on the homepage, each representing a different navigation path for a different kind of visitor (called *audience track*) containing all information/functionality relevant for that kind of visitor.

Although this design philosophy significantly reduces the amount of information the visitor needs to plough through, it can also be a cause of annoyance if the user cannot find the information he is looking for in the chosen audience track, or if his track contains information of no interest to him. As it is more difficult for web designers[1] to

---

[1] Target audiences for web sites are often more difficult to access and perform standard requirements engineering techniques (e.g. questionnaires) upon, compared to classical prospectus users of a standard application.

assess the exact requirements relevant or irrelevant for a certain user, it is well possible that some information fulfilling a requirement is missing in the audience track for a certain visitor (but present in another), or other information is put wrongly in that track. In this paper, we will tackle this problem by describing how to identify such missing or superfluous information for a certain target audience, and how to correct for it, by adapting the structure and navigation in the website. The framework for this work is the Web Site Design Method (WSDM) [6], an audience driven design method.

Web design methods that have support for adaptation include WebML[3], Hera[7], UWE[9] and OOH[8]. However, these methods focus either on personalization (e.g. towards different devices, content personalization), or handle a bottom-up approach to constructing pages from basic concepts. As far as the authors are aware of, there is no other research combining design support for altering an existing structure based upon user access, to improve the structure of the site towards all users.

2   Identifying Missing or Superfluous Information

Three steps are involved in detecting superfluous or missing information in a certain audience track:

1. **Determine to which audience class the current user belongs**
    Due to audience driven design, the user has to choose an audience track with his first click, thus choosing to which audience class he belongs.

2. **Determine if and which information the user visits, both within and outside his audience track.**
    As we want to keep track of which information is visited outside a particular audience track, and relate this information to the frequency of visits inside the track, we cannot just store the (total) number of accesses to every piece of information (modeled as a *chunk* in WSDM). Instead, we need to store the number of visits to each piece of information relative to each audience class. This data can be conveniently stored in a matrix, which we will call the *information access matrix*. Rows of the matrix represent the information (chunks) that represents the different elementary requirements, while the columns represent the different audience classes. Over time, the matrix contains a good summary of the amount of accesses to the different chunks for each audience class.

3. **Analyze the accumulated data to determine if superfluous or missing information for a certain audience track is detected.**
    We fall back on known statistical techniques: we consider the problem of detecting missing information as the problem of deciding whether a certain value (i.e. the amount of accesses to *foreign information,* information outside an audience track) *fits well* in a given sample data set (i.e. the set of amounts of accesses to *native* information, information within the audience track). Although different statistical techniques can be used, we have chosen in this paper to use median (which is robust) as a measure of central tendency combined with median absolute deviation as a measure of spread. As most values (see [4] for a more exact estimate) of a dataset

probably lie within the distance of the spread from the middle value, we can use (median – MAD), calculated over the set of accesses to native information as a threshold value above which accesses to foreign information is considered relevant.

We consider the problem of detecting superfluous information as the problem of detecting outliers [1] (i.e. the amount of accesses of possible superfluous information) within a given dataset (i.e. the set of amount of accesses to *native* information). Although different statistical techniques can be used, we have chosen in this paper to use a double strategy to find *low* values: find values that lay both *far* below the middle value of the given dataset and *far* from their bigger neighbor. To determine which point lies *far* from the middle value, we again calculate the median and MAD for the given dataset, and test which values of the dataset lay below the threshold (median – MAD). To determine which values are far from their neighbor, we take the ordered dataset, and calculate the distances between each 2 consecutive values. These distances give us a new dataset, for which we calculate mean and standard deviation (= std). Calculating the interval [(mean–std) (mean+std)] gives us the range in which most of the distances (i.e. 50% for normally distributed data) lay. Distances above the (mean+std) are thus relatively big, compared to the other distances, and we have identified two points that are relatively far from each other.

3   Clarifying example

To clarify this method, let's consider a real life example of a website (partly) built according to an audience driven design philosophy: the NASA web site (http://www.nasa.gov/). The information access matrix for these tracks is shown in figure 1. For simplicity, we consider only the audience classes "Media & Press" (with *native* information *inf1* to *inf6*) and "Educators" (with native information *inf7* to *inf12*). As explained, each cell in the matrix denotes the number of visits to some information by a certain audience class. As we were unable to obtain the *real* access information for the NASA website, we have used for this example fictitious data.

		Media & Press	Educators
Inf.1	Press Release Archive	52	40
Inf.2	Press Contacts	49	4
Inf.3	Press Kits	31	10
Inf.4	Fact Sheets	16	5
Inf.5	Speeches	40	5
Inf.6	Images	38	12
Inf.7	Contacts for educators	0	56
Inf.8	Professional development	5	50
Inf.9	Student opportunities	1	30
Inf.10	Fellowships and grants	0	10
Inf.11	Teaching Internet Resources	3	20
Inf.12	Teaching Multimedia Resources	2	15

Figure 1. Information Access Matrix

Let's now analyze the accesses to the native information of the Educators audience track (inf7 … inf12), and determine if accesses to foreign information (inf1 ... inf6) were significant compared to these accesses:

Data set (ordered): 10 15 20 30 50 56
Median: 25 ; MAD: 12.5 ; Threshold: 25 – 12.5 = 12.5
Detected foreign requirements: Press Release Archive (40 accesses)

We can thus conclude that this information should somehow be included in the Educator track (how and where this information is added is the subject of the next section). Let's now consider the calculations to identify possible superfluous information in the "Media & Press" audience track.

Data set (ordered): **16** 31 38 40 49 52   Data set of distances: **15** 7 2 9 3
Median: 39; MAD: 9   Mean: 7.2 ; Standard deviation: 4.7
Lower limit: 39 – 9 = 30   Upper limit: 7.2 + 4.7 = 11.9

As the first element ('Fact Sheets', 16 accesses) of the original dataset lies below the corresponding lower limit (30), and the distance between the first and the second element (15) lies above the corresponding threshold (11.9), we can conclude that 'Fact Sheets' is detected as (possibly) superfluous for Media & Press audience class.

4   Correcting Missing or Superfluous Information

Having identified missing or superfluous information in a certain audience track, the structure of the web site can be adapted (automatically) to correct the detected deficiencies. To specify this possible adaptation at design time, we use the Adaptation Specification Language (ASL) [2] defined over the navigation model of WSDM.

Possible actions taken upon detection of missing information within an audience track varies from duplicating the information in the place it is missing, provide a link to the existing information, or totally re-arrange the structure of the site. The approach shown in this paper consists of duplicating the detected nodes, and offering a link to the information at the root of the audience track:

Foreach *AudienceClass* **in** Website
  Foreach *node* **not in** NodesOfAudienceTrack(*AudienceClass*):
    if *node* **in** *MissingInformation(AudienceClass)*
    then addLink(root(Audience*Track(AudienceClass)*, duplicate(*node*))

Information identified as superfluous in a certain audience track does not necessarily need to be removed. Although visited only few times, it might still be valuable for (a small amount of) visitors. Automatic adaptation is possible (for example, apply link sorting), but more tricky and lengthy to describe; due to space restrictions we cannot describe it here. For now, we consider the detection of superfluous information rather as an alert to the web master, than something that requires (automatic) adaptation.

6 Conclusions & Further Work

In this paper, in the context of an audience driven design method for web sites, WSDM, we provide a mechanism for adaptively correcting the structure of the web site. By observing the user browsing behavior and the use of statistical techniques, missing or superfluous information in a particular audience track are automatically detected. At design time, the designer can specify (using the Adaptive Specification Language) the adaptive actions that should be taken in case such situations are detected at run time. By doing so, the structure of the web site will be better tailored to the needs of the different audience classes.

Future work includes accommodating the Adaptation Specification Language to make it possible to acquire all necessary information (e.g. creating and updating the information access matrix). Further research in analyzing the data from the matrix is being performed. In particular, detecting correlations between (the same) missing or superfluous information in different audience classes, and a cleverer way to adapt the structure of the site accordingly. Different adaptation strategies upon detection of missing/superfluous information are also a way to continue the research.

References

1. Barnett, V. Lewis, T.: Outliers in Statistical Data. Wiley 3rd Edition (1995)
2. Casteleyn, S., De Troyer, O., Brockmans, S.: "Design Time Support for Adaptive Behaviour in Web Sites", In Proceedings of the 18th ACM Symposium on Applied Computing, Melbourne, USA (2003)
3. Ceri S., Fraternali P., and Bongio A: "Web Modeling Language (WebML): a modeling language for designing Web sites", In WWW9 Conference, First ICSE Workshop on Web Engineering, International Conference on Software Engineering (2000)
4. DeGroot, M..H.: Probability and Statistics. Addison-Wesley, Reading, Massachusetts (1989)
5. De Troyer, O.: "Audience-driven web design", In Information modelling in the new millennium, Eds. Matt Rossi & Keng Siau, IDEA GroupPublishing, ISBN 1-878289-77-2 (2001)
6. De Troyer, O. ,Leune, C.: "WSDM: A User-Centered Design Method for Web Sites", In Computer Networks and ISDN systems, Proceedings of the 7th International World Wide Web Conference, Elsevier, pp. 85 – 94 (1998)
7. Frasincar, F., Houben G., and Vdovjak R.: Specification Framework for Engineering Adaptive Web Applications, In Proceedings of the World Wide Web Conference, Honolulu, Hawaii, USA (2002)
8. Garrigós, I., Gómez, J. and Cachero, C.: "Modelling Dynamic Personalization in Web Applications", Third International Conference on Web Engineering (ICWE'03), LNCS 2722, pages 472-475. Springer-Verlag Berlin Heidelberg (2003)
9. Koch, N.: "Software Engineering for Adaptive Hypermedia Systems, Reference Model, Modeling Techniques and Development Process", PhD Thesis, Verlag UNI-DRUCK, ISBN 3-87821-318-2 (2001)
10. Nielsen, J.: Finding usability problems through heuristic evaluation. Proceedings of the SIGCHI conference on Human Factors and Computer Systems. Monterey, California, United States Pages: 373 – 380. ISBN:0-89791-513-5 (1992)