ABSTRACT

This paper describes a neurofuzzy-based approach to automatically classify publicly accessible World Wide Web sites. The suggested methodology appears most relevant investigating corporate Web sites for advertising and customer support purposes but is equally valuable for analysing the hypertext structures of educational or non-profit organisations. Future efforts will focus on extending the framework from pre- and after-sales functions to customizable electronic transactions.

A taxonomy is proposed based on "hard" criteria for classifying Web sites. Research methods and phases for a validated model are outlined and the most relevant attributes of such a system are defined. The described Web site analysis tool supports the automated data gathering and assures the necessary "critical mass" of training cases. The gathered data and the Web site taxonomy are fed into a neurofuzzy system which uses supervised learning in the assessment tool and unsupervised learning (fuzzy clustering) for validation.

1. INTRODUCTION AND RESEARCH OBJECTIVES

The principal idea behind this research is to employ an autonomous software tool to capture the characteristics of a specific Web site. Many attributes of Web pages and sites can be determined and measured automatically by programs and subsequently stored into a database. The utilisation of dedicated software agents (i.e., a "robot") to examine Web sites is more efficient and immune against inter- and intrapersonal variances than human assessment. Thus the inclusion of thousands of Web sites becomes feasible compared to tens (Selz et al., 1998) or hundreds (http://www.excite.com) in previous efforts. Naturally, these advantages come at the expense of sacrificing non-quantifiable information and recipient-dependent attributes. It would require human intervention to correctly identify and assess these "soft" attributes.

The ultimate goal is to develop a consistent classification framework of publicly accessible Web sites based on a software tool determining the most important variables through processing hypertext documents available via the Internet. Both the classification framework and the software tool imply the definition of measurable criteria. These criteria will be investigated and preprocessed by the tool for each Web site separately and serve as input for the assessment model. Possible applications of such a tool fall into three areas:

1. Static or snapshot analysis: Analysis of a (large) number of Web sites at a given time allows comparison of individual criteria to means, variances and other statistical parameters and classification of Web sites into categories. When clustered into sub-groups - for example by industry, by company size, by or experience in online business - comparisons between these sub-groups and between sub-groups and the whole population become feasible.

2. Dynamic or longitudinal analysis: Changes to a set of Web sites can be measured and reported. Development over time, recent trends and drastic changes can be monitored for specific industries or other sub-groups.
3. Comparative analysis: At company level the results will enable marketing and IT managers to assess competitors' Web sites and critically compare them with their own achievements.

2. THEORETICAL FOUNDATIONS OF WEB SITE EVALUATION

Elements of classification and evaluation of World Wide Web resources can be applied to a variety of situations. These approaches basically employ two methods of evaluation: manual versus automated assessment. In most cases, manual assessment relies on the judgement of individual analysts evaluating a certain Web site. Even if a group of experts is involved, the assessment takes place with varying degrees of process and data structure in capturing the required information. Rigorous structure is assured when evaluation is done automatically with software tools. However, most often this advantage of consistency leads to less relevance for some specific domains of knowledge.

2.1 Manual Assessment

Developing quality criteria is the first step towards any classification framework based on empirical evidence (in contrast to conceptual frameworks). Assessment methodologies for reviewing Web sites are a prerequisite for validation of such frameworks with the target group ranging from the reviewers themselves to the international Internet community. Documented approaches can be found in the following areas:

1. Awards and prices: Ever since the commercialisation of the World Wide Web, net awards and commercial ratings gained in popularity - e.g., Point Communications' Top 5% of All Websites Award, Magellan 4 Star Sites, Lycos Top 5%, WWW Associates Top Ten, etc. The evaluation process relies on either some sort of selection panel (e.g., CommerceNet Bestows Awards for Electronic Commerce Excellence; http://www.commerce.net/news/press/19981026.html) or public voting (compare the Australian Internet Awards; http://www.webawards.info.au/webawards/awards.htm). More sophisticated approaches, such as SurveySite (http://www.surveysite.com), employ client-side applications to capture user opinions and consumer perceptions.

2. World Wide Web search and index pages: URL collections on the World Wide Web like Yahoo! (http://www.yahoo.com) represent classical examples of Web site classification. Various criteria are employed ranging from subject to geographical indices. Occasionally, the user value of such collections is further improved by adding detailed reviews of indexed sites - e.g., Excite (http://www.excite.com).

3. (Information) Quality assessment of Internet resources: Checklists and guidelines have been developed to allow even inexperienced users assessing the quality and reliability of various Internet resources - for a list of online resources compare Elliott (1998).

4. Assessing the business value of a Web Information System: As far as the commercial use of Web sites is concerned, this last category proves most important. Selz et al. (1997; 1998) propose a "Web Assessment Model" to identify and evaluate successful Electronic Commerce applications. While the resulting model offers a very detailed insight into these solutions, the time-consuming method requires access to frequently not available company information.

2.2 Automated Assessment

Software developers and the Internet community have introduced a number of automated systems to check everything from the basic questions of syntactical correctness and browser compatibility of HTML code, or spelling to problems of higher complexity like content-rating of a Web site or the quality of its representation. These developments can roughly be divided into two categories completely independent from each other:

1. Quality checks of Web pages: These tools work mainly as a help for Web site designers to ensure the quality of their work (e.g., W3C HTML Validation Service, http://validator.w3.org) and frequently are integrated into the design tools themselves. The output of more sophisticated online tools - e.g., Web Site Garage (http://www.websitemaintenance.com), Fritz-Service (http://www.fritz-service.com), or Bobby (http://www.cast.org/bobby) - includes multi-dimensional ratings of the investigated Web sites.
2. Content determination of Web pages: Metadata labels (e.g., Platform for Internet Content Selection; http://www.w3.org/PICS/) appear to be the most promising technology in this area. Metadata enhances retrieving information through search engines and may help avoiding undesired content.

2.3 Previous Classification Approaches

Frameworks for Web site classification have been introduced previously. The majority of these efforts are driven by scholarly desire for typology. Many of them have a strong foundation in advertising and are utilised as frameworks for strategic decision making and for defining the marketing goals of online businesses. However, a clear distinction between identified categories is very often lacking as most of these classifications are based more on anecdotal than empirical evidence at significant level. For this research a Web site taxonomy has been developed based on the classification criteria mentioned above and on advancing and comparing previously introduced classification schemes. The proposed Web site taxonomy will then be validated through a neurofuzzy model (see section 3 and Figure 1 for a detailed description of the research efforts).

Hoffman et al. (1995) identify six distinct categories for commercial Web sites from an integrated marketing point of view. Their classification is based on a functional typology and an integrated and complete approach would include all types. The following types were identified:

1. **Online storefront** offers direct sales, electronic catalogues and/or other forms of electronic commerce.
2. **Internet presence** can take the form of flat ads (single Web page), image sites (emotional consumer appeal) or detailed information sites.
3. **Content sites** are characterised by their funding model, which can be either fee-based, sponsored, or a searchable database.
4. **Electronic Malls** typically feature a collection of online storefronts.
5. **Incentive Sites** make effective use of the specific technical opportunities of the Web for advertising purposes.
6. **Search agents** identify other Web sites.

A generalised commercial Web site classification with similar intentions, but a mutually exclusive typology, has been given by Hansen et al. (1996). It comprises two dimensions, content and interactivity. Five distinct types of Web sites - Electronic Business Card, Advertainment, Electronic Brochure, Electronic Catalogue and Web-Service/Shop - are defined by putting them into the context of these dimensions. Similar schemes can be found for industry-specific Web site classification. Such models, for example, have been applied in the retail banking industry for building reference models (Bauer, 1997), observation of industry evolution (Mahler et al., 1996), and commercial Web site analysis (e.g. Booz, Allen and Hamilton, 1998; Unisys, 1998).

With regard to Web-based information system adaptability, Scharl and Brandtweiner (1998) offer a more technology-oriented classification scheme and distinguish between five different levels of Web-based customization (also compare Hansen and Tesar, 1996). They present their scheme in conformity with the ecology-oriented framework of James F. Moore (1993, 1997). Assuming that the development of electronic markets follows the four distinct stages identified by Moore - Birth, Expansion, Authority, and Renewal (or Death) - designers of Web information systems currently find themselves at the beginning of the expansion stage which is mainly characterized by the creation of a world-wide critical mass of customers in electronic markets.
3. RESEARCH PLAN AND METHODOLOGY

To satisfy the requirements of (qualitative) model building and (quantitative) model validation a multi-methodological approach has been chosen. Research methodologies applied include secondary data analysis, exploratory case studies, prototyping for proof-of-concept, and statistical testing for significance. After the initial identification of classification criteria a hypothetical Web site taxonomy has been developed. An analysis tool is currently under development which will be able to automatically gather the classification criteria explicitly stated in Table 1 (with the exception of password-protected segments of Web sites requiring manual registration). Future versions of this tool will include a component for exploratory textual analysis based on linguistic units like words or lemmas.

By means of linear step-wise regression analysis, the program will filter out the most important independent variables (= input vector) from this set of attributes for inclusion in the final model. Subsequently, the assessment tool will use the adaptive neurofuzzy architecture of Figure 2 to determine the appropriate category (= output vector) for any given Web site. The optimisation of the neural component integrated into this assessment tool will be done by supervised learning using a standard backpropagation algorithm for specifying rules and membership functions of the fuzzy inference engine (FAM inference; Inform, 1997). For validation purposes, a second neurofuzzy system will be employed to cluster the available data (= unsupervised learning), independent of the taxonomy specified in advance. By comparing the categories of the deductive approach (taxonomy) with the results of the inductive approach (generated fuzzy clusters) we expect to gain valuable insights as far as the appropriateness of current models and concepts for the evaluation of Web sites are concerned.

The classification criteria are critical for the success of the assessment tool. Considering technical feasibility, the sub-criteria identified in Table 1 evolved from analysis of previous approaches and - to some extent - from exploratory case studies. The sub-criteria are grouped into five categories, which are characterised by varying degrees of measurability. It is, for example, much harder to capture "content" in quantitative variables than the level of (transaction-oriented) security. However, on sub-criteria level each item is matched by at least one quantifiable variable. The quality of the classification framework represents a function of the quality of sub-criteria definition.
Table 1. Classification criteria
(# = number of; Nom = nominal; Ord = ordinal; Num = numeric)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Sub-Criteria</th>
<th>Variable</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Wealth of Information</td>
<td>KB of Text</td>
<td>Num</td>
</tr>
<tr>
<td></td>
<td>Size</td>
<td># Pages</td>
<td>Num</td>
</tr>
<tr>
<td></td>
<td>Additional Information</td>
<td># Links to non-HTML Files (.pdf, .doc, .rtf, .ps, etc.)</td>
<td>Num</td>
</tr>
<tr>
<td>Interactivity</td>
<td>CGI</td>
<td># Online Forms</td>
<td>Num</td>
</tr>
<tr>
<td></td>
<td>CGI-Usage</td>
<td>Online Form Fields</td>
<td>Ord</td>
</tr>
<tr>
<td></td>
<td>Client-Side Information</td>
<td>Cookies</td>
<td>Ord</td>
</tr>
<tr>
<td></td>
<td>Internet Programming</td>
<td>Java-Applets, ActiveX</td>
<td>Nom</td>
</tr>
<tr>
<td></td>
<td>E-Mail Interaction</td>
<td># Differing &quot;Mailto:~-Links</td>
<td>Num</td>
</tr>
<tr>
<td>Design</td>
<td>Size</td>
<td># Pages, KB</td>
<td>Num</td>
</tr>
<tr>
<td></td>
<td>Dialogs / Enhanced</td>
<td>JavaScript</td>
<td>Nom</td>
</tr>
<tr>
<td></td>
<td>Graphics</td>
<td># Images, KB</td>
<td>Num</td>
</tr>
<tr>
<td></td>
<td>Navigational Structure</td>
<td>Use of Frames</td>
<td>Nom</td>
</tr>
<tr>
<td></td>
<td>Connection to Other Sites</td>
<td># External Links</td>
<td>Num</td>
</tr>
<tr>
<td></td>
<td>Navigation</td>
<td># Internal Links per Page</td>
<td>Num</td>
</tr>
<tr>
<td>Web Site Management</td>
<td>Content Maintenance</td>
<td>Expired Dates</td>
<td>Ord</td>
</tr>
<tr>
<td></td>
<td>Structural Integrity</td>
<td>% Broken Links</td>
<td>Num</td>
</tr>
<tr>
<td></td>
<td>Web Site Tool Integration</td>
<td>Metatags</td>
<td>Ord</td>
</tr>
<tr>
<td>Security</td>
<td>Simple Password Protection</td>
<td>HTTP-Authenticate</td>
<td>Nom</td>
</tr>
<tr>
<td></td>
<td>Encryption</td>
<td>SSL, SHTML</td>
<td>Nom</td>
</tr>
</tbody>
</table>

Figure 2. Neurofuzzy component created with FuzzyTech 5.01 Professional (Inform, 1997)

4. CONCLUSION

With the automated Web site evaluation approach described in this paper it becomes feasible to assess thousands of Web Sites at a given time without human intervention. Repeated evaluations support the tracking of dynamic trends within certain industries. With fuzzy clustering as the inductive validation method, the common mistake of arbitrary definitions during the initial specification of categories is avoided and will lead to an assessment model more appropriate for real-world Web information systems. While fuzzy modelling maintains the transparency of the system's internal calculations, the neural component ensures maximum flexibility and continued optimisation, eliminating the need for predefined and often questionable mathematical relationship.
5. REFERENCES

BAUER, C. (1997). Masseninformationssysteme für Banken: Referenzmodelle unter besonderer Berück- 
sichtigung des World Wide Web (WWW). Doctoral Thesis, Vienna University of Economics and 
Business Administration, Vienna.

New York.


HANSEN, H.R. and TESAR, M.F. (1996): "Die Integration von Masseninformationssystemen in die be-
triebliche Informationsverarbeitung", "Data Warehouse", Gerhard Mercator University Duisburg. n.a.


488-492.

(3), 75-85.

MOORE, J.F. (1997). The death of competition - leadership and strategy in the age of business ecosys-


ment of Successful Electronic Commerce Applications. 31st Hawai‘i International Conference on 
System Sciences (HICSS-31). Volume IV: Internet and the Digital Economy. IEEE Computer Society, 
Los Alamitos, 222-231.