Quantitive evaluation of Web site content and structure

Christian Bauer and Arno Scharl

The authors

Christian Bauer (bauerc@cbs.curtin.edu.au) is at Electronic Commerce Network, Curtin University of Technology, Perth, Western Australia.

Arno Scharl (scharl@wu-wien.ac.at) is at the Information Systems Department, Vienna University of Economics & BA, Vienna, Austria.

Keywords

Web site classification, Neural networks, Non-profit organizations

Abstract

Describes an approach automatically to classify and evaluate publicly accessible World Wide Web sites. The suggested methodology is equally valuable for analyzing content and hypertext structures of commercial, educational and non-profit organizations. Outlines a research methodology for model building and validation and defines the most relevant attributes of such a process. A set of operational criteria for classifying Web sites is developed. The introduced software tool supports the automated gathering of these parameters, and thereby assures the necessary "critical mass" of empirical data. Based on the preprocessed information, a multi-methodological approach is chosen that comprises statistical clustering, textual analysis, supervised and non-supervised neural networks and manual classification for validation purposes.

Electronic access

The current issue and full text archive of this journal is available at

http://www.emerald-library.com

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 1 · 2000 · pp. 31–43
© MCB University Press · ISSN 1066-2243

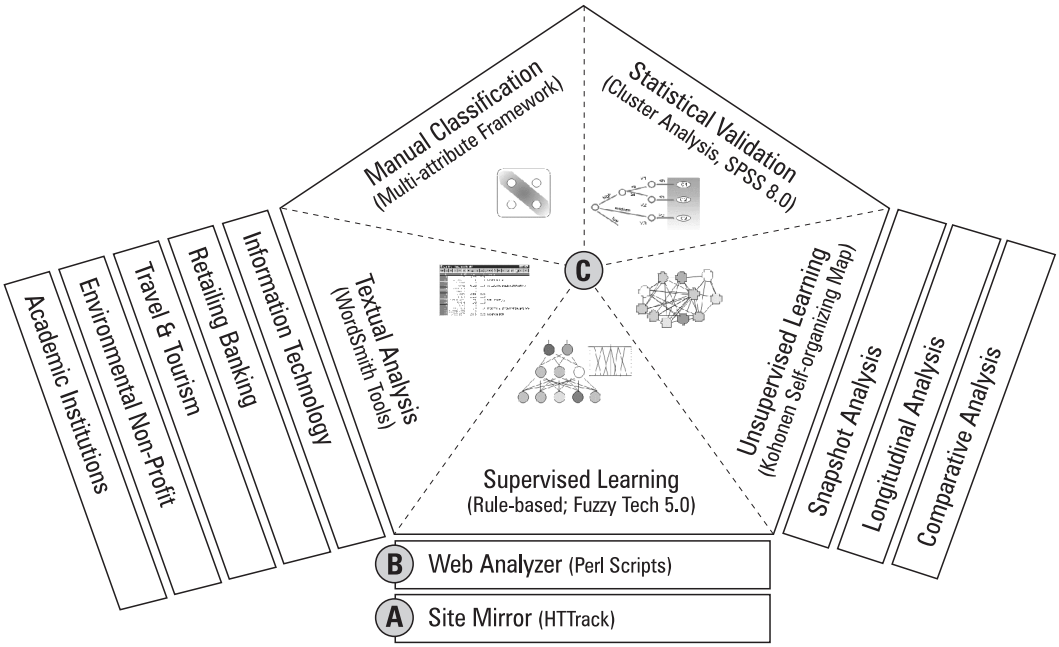
Introduction and analytical objectives

The principal idea behind this paper is the use of autonomous software tools to capture the characteristics of commercial Web information systems, determine their specific importance, and store them in a central data repository. The utilization of dedicated software agents to examine Web sites is more efficient and immune against intra- and interpersonal variances than human evaluation. Thus, the inclusion of thousands of systems becomes feasible, compared to samples limited to tens or hundreds in previous efforts (Bucy *et al.*, 1999; Selz and Schubert, 1997; Witherspoon, 1999). Naturally, these advantages come at the expense of sacrificing non-quantifiable, frequently recipient-dependent information. These previous efforts conducted manual Web site assessment, and therefore did not have access to the necessary resources to cover more significant samples. Schubert and Selz (1999), for example, developed an on-line survey tool with an extensive list of evaluation criteria for Web assessment. And while their results include a plethora of valuable information, their data collection was limited to around 70 assessments for several Web sites.

The ultimate goal is to develop a consistent analysis and evaluation framework of publicly accessible hypertext structures. However, we have to be cognizant of the relevant attributes to be able to evaluate and group them (Tesch, 1990). Thus both the framework and the software prototype imply the definition of measurable, operational criteria. These criteria will be investigated and preprocessed by the tool for each Web site separately. Possible applications of such a tool fall into three areas (see Figure 1):

- *Snapshot analysis (static)*. Analysis of a large number of Web sites at a given time allows comparison of individual criteria (e.g. means, variances, or other statistical parameters), and the clustering of sites by business sector, company size, or experience with electronic transactions. From an academic point of view, the discerning of similarities helps to refine the discriminative power of the established categories (Tesch, 1990).

Figure 1 Research methodology



- *Longitudinal analysis (dynamic)*. A defined set of Web sites can be documented and analyzed over a longer period of time, i.e. a series of snapshot analyses are conducted on the same sample. Trends, cyclical variations, turning points, and marginal changes are monitored for specific industries and other sub-categories. These complex developments can hardly be observed and quantified in a traditional way. Nevertheless, both industry and academic research require the generated indicators.
- *Comparative analysis*. In the case of real-world organizations that have to assess alternative investment opportunities, the analysis of their own business sector and competitors' efforts provides decision makers with reliable benchmark data about relevant differences, and the relative performance of their own sites. Benchmarks indicating weaknesses should trigger and guide the redevelopment and optimization of deployed systems. Especially this third category provides companies with an important source of information for maximizing their customer delivered value.

Generally, we must distinguish between manual and automated approaches for analyzing (positivistic aspect) and evaluating (normative aspect) on-line resources. In most cases, manual evaluation relies on the judgments of individual analysts on certain Web sites. Even if a group of experts is involved, the evaluation takes place with varying degrees of subjectivity regarding the process itself and the captured data structures. In addition to that, the highly dynamic nature of the World Wide Web necessitates rapid collection of data, limiting manual approaches to consider much beyond the system's introductory homepage (McMillan, 1999). Speed, rigorous structure and abundance of data are assured when evaluation is done automatically with software tools. However, this advantage of consistency may also lead to reduced relevance for specific domains of knowledge.

Manual versus automated evaluation

Several frameworks for Web classification, both formal and informal, previously have been proposed before. Most of these efforts are driven by scholarly desire for typology, viewed not only as intellectual tools for organizing data segments, but also as significant research results. They usually have a strong foundation in advertising and are utilized as guidelines for

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 1 · 2000 · 31–43

strategic decision making and for defining the marketing goals of on-line businesses. Hoffman *et al.* (1997), for example, identify six distinct categories for commercial Web sites based on a functional typology. On-line storefronts offer direct sales, electronic catalogs, and substantial customer support. An Internet presence can take the form of flat ads (single documents), image sites (emotional consumer appeal), or detailed information sites. Content sites are characterized by their funding model, which can be either fee-based, sponsored, or a searchable database. Electronic malls typically feature a collection of on-line storefronts. Incentive sites make effective use of the specific technical opportunities of the Web for advertising purposes. Services belonging to the last category, search agents, identify other Web information systems.

Hansen (1996) presents a generalized commercial classification with similar intentions, but a mutually exclusive typology. It comprises two dimensions (content and interactivity) and defines five distinct types of Web sites (electronic business cards, advertainments, electronic brochures, electronic catalogs and Web services) in the context of these dimensions. Similar schemes can be found for industry-specific Web site classification. They have been applied in the information technology and retail banking industries for building reference models (Scharl, 1997; Bauer, 1998), observation of industry evolution (Mahler and Göbel, 1996), and commercial Web site analysis (Booz-Allen and Hamilton, 1999; Psoinos and Smithson, 1999). With regard to system adaptivity, 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). However, many of these classification frameworks lack a clear and consistent distinction between identified categories as many of these approaches are based more on anecdotal than empirical evidence.

The methodology presented in the following is based on the classification criteria listed in Table I, and tries to advance previously introduced classification schemes by validating them with a neural network (Kohonen self-organizing map). Specifying these criteria is the first step towards any evaluation framework based on empirical evidence (in contrast to conceptual frameworks). Considering technical feasibility, the three categories of Table I, which are characterized by varying degrees of measurability, evolved from reviewing previous approaches and exploratory case studies. Even basic attributes such as the total number of documents may serve as good indicators for the Web genre a particular site belongs to (see Table II).

Table I Domain-independent, automatically collected classification criteria

Criteria	Variables
Content	No. documents [total]
	Kbytes downloaded [total/text only]
	No. file types [distinct extensions]
	No. images [total/distinct]
Interactivity	No. forms [total/distinct/fields]
	No. documents w/JavaScript [total]
	No. Java applets [total/distinct]
	No. MailTo-links [total/distinct]
Navigation	Frames [yes/no]
	No. internal links [total/distinct/broken]
	No. external links [total/distinct/broken]
	No. anchors [total]
	No. links to anchors [within/between documents]

To date, much of the analysis about Web content and form has been qualitative or proprietary to industry and only partially released (Bucy *et al.*, 1999). In order to overcome the methodological limitations of subjective impressions, anecdotal evidence, or convenience samples, a number of automated systems have been introduced as well. All empirical approaches have the attention to establish some differentiation between Web sites, from the basic questions of spelling, syntactical correctness, and browser compatibility of HTML (Hypertext Markup Language) code to problems of higher complexity such as Web sites content-rating or the perceived quality of its presentation. Documented approaches of empirical Web site classifications can be found in the following areas:

- *Awards and prices*. With the commercialization of the World Wide Web, awards and commercial ratings gained in

Table II Approximate sizes of Web genres (Shneiderman, 1997)

Total no. of documents	Example genres
1-10	Personal bio, restaurant review, project summary, course outline
5-50	Scientific paper, photo portfolio/exhibit, conference program, organization overview
50-500	Book or manual, city guide/tour, corporate annual report, product catalog/advertisement
500-5,000	Photo library, museum tour, technical reports, music/film database
5,000-50,000	University guide, newspaper/magazine
50,000-500,000	Telephone directory, airline schedule
> 500,000	Congressional digest, digital library
> 5,000,000	Library of Congress, NASA archives

popularity, e.g. Point Communications' top 5 per cent of all Web sites award, Magellan four-star sites, Lycos top 5 per cent, 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)[1] or public voting (compare the Australian Internet Awards; http://www.webawards.info.au/). More sophisticated approaches, such as SurveySite[2], employ client-side applications to capture user opinions and consumer perceptions.

- *Content determination*. Semantic labels, e.g. Platform for Internet Content Selection (PICS)[3], appear to be the most promising technology in this area, especially with regard to semantic markup languages like XML. They enhance information retrieval and avoid undesired content. Search engines, directories and meta indices like Yahoo![4] represent classical examples of Web site classification and content determination. Various criteria are employed including subject and geographical area. Occasionally, the perceived value of such collections is further improved by adding detailed reviews of indexed systems, e.g. Excite[5].
- *Quality assessment of Internet resources*. Checklists and guidelines have been developed to allow even inexperienced users to assess the quality and reliability of various Internet resources[6]. Some of these approaches combine manual and automated gathering of required data, facing rigid limitations regarding feasible sample sizes and limiting the objectivity of the approach. The Web site quality evaluation method (QEM) proposed by Olsina *et al.* (1999), for example, uses the hierarchical system of attributes listed in Table III to assess the artifact quality of academic Web sites. Although quite impressive at first sight, too many attributes raise some subtle problems of computational nature, as the predictive power of statistical algorithms and methods based on neural network tends to suffer from multi-dimensional input vectors of such high complexity. The desired generalization can often be better achieved using only a few, but highly relevant, features of the sample to be analyzed. The domain-dependency of Olsina *et al.*'s (1999) quality requirement tree further limits the potential scope of its application. More general tools help Web designers assess the quality of their work (e.g. W3C HTML validation service)[7] and are frequently integrated into the underlying design architecture. The output of more sophisticated online tools such as the Web Site Garage[8], Fritz-Service[9] or Bobby[10] includes multi-dimensional ratings of investigated Web sites.
- *Assessing the business value of Web information systems*. For commercial scenarios, this last category proves most important. Selz and Schubert (1997) propose a specific Web assessment model to identify and evaluate successful commercial applications. While the resulting model offers a detailed analysis of these solutions, the time-consuming method requires access to company information, which frequently is not available. The e-audit methodology used for the *1999 Worldwide Web 100 Survey* of the London School of Economics (Psoinos and Smithson, 1999) represents a more pragmatic approach. The authors of this report audited 120 Web sites across eight business sectors, with most survey participants drawn from the 1998 *Fortune* 500 global rankings. Mimicking all the

Table III Domain-dependent, manually collected classification criteria (Olsina *et al.*, 1999)

Functionality	Usability	Efficiency	Site reliability
Searching and retrieving issues	*Global site understandability*	*Performance*	*Link errors*
• Web site search mechanisms:	• Global organization scheme:	• Static page size	• Dangling links
– scoped search (people, course, academic unit)	– site map		• Invalid links
– global search	– table of contents	*Accessibility*	• Unimplemented links
• Retrieve mechanisms:	– alphabetical index	• Information accessibility:	
– level of retrieving customization	• Quality of labeling system	– support for text-only version	*Miscellaneous errors or drawbacks*
– Level of retrieving feedback	• Student-oriented guided tour	– image title	• Deficiencies or absent features due to different browsers
	• Image map (campus/buildings)	– global readability	• Deficiencies or unexpected results (e.g. non-trapped search errors, frame problems, etc.) independent of browsers
Navigation and browsing issues	*On-line feedback and help features*	• Window accessibility:	• Dead-end Web nodes
• Navigability:	• Quality of help features:	– no. pages regarding frames	• Destination nodes (unexpectedly) under construction
– indicator of path	– explanatory help	– non-frame version	
– label of current position	– search help		
– average of links per page	• Web site last update indicator:		
• Navigational control objects:	– global		
– contextual controls permanence	– scoped (per sub-site or page)		
– contextual controls stability	• Addresses directory:		
– vertical scrolling	– e-mail directory		
– horizontal scrolling	– phone/fax directory		
• Navigational prediction:	– post mail directory		
– link title (link with explanatory help)	• FAQ feature		
– quality of link phrase	• On-line feedback:		
	– questionnaire feature		
Student-oriented domain features	– guest book		
• Content relevancy:	– comments		
– academic unit information (index, sub-sites)	*Interface and aesthetic features*		
– enrollment information (entry requirement information, form fill, download)	• Cohesiveness by grouping main control objects		
– degree information (index, description, course offering)	• Presentation permanence and stability of main controls:		
– course description (syllabus, comments, scheduling)	– direct controls permanence		
– student services information (index, healthcare, sport, scholarship, housing, culture)	– indirect controls permanence		
– academic infrastructure information (laboratory, library, research results)	– stability		
• On-line services:	• Style issues:		
– grade/fees on-line information	– link color style uniformity		
– Web service	– global style uniformity		
– FTP service	– global style guide		
– newsgroup service	• Aesthetic preference		
	Miscellaneous features		
	• Foreign language support		
	• What's new feature		
	• Screen resolution indicator		

stages of real on-line transactions, the business transaction cycle is divided into seven dimensions, which are then assessed separately:
  - company information;
  - advertising and promotion;
  - product information;
  - order;
  - settlement;
  - after-sales service; and
  - ease of use/innovation.
- *Application benchmarks*. Several academic and commercial efforts aim at providing standardized methods to estimate technical parameters such as system response times under various loads, throughput, or scalability. Jutla *et al.* (1999) mention the following examples: Webperf from the Standard Performance Evaluation Corporation (SPEC)[11], WebStone, Benchmark Factory 97, the upcoming TPC-W from the Transaction Processing Performance Council (TPC)[12] and WebEC, the latter being developed by the Jutla *et al.* (1999) themselves.

Methodology

To satisfy the requirements of a new, multi-disciplinary area combining quantitative and qualitative model building and validation, a multi-methodological approach has been adopted. According to Galliers' (1993) classification of information systems research, research methodologies applied include interpretive data analysis, exploratory case studies, and prototyping for theorem proof. It also covers the four research strategies for a multi-methodological approach to information systems research as proposed by Nunamaker *et al.* (1990/91):
(1) theory building;
(2) experimentation;
(3) observation; and
(4) systems development.

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 gather automatically the classification criteria stated in Table I (with the exception of password-protected segments of Web sites that require manual registration).

Research objectives

The three application areas (snapshot, longitudinal, and comparative analysis) identified in the introduction rely on successfully answering the following research questions:
(1) What information about Web information systems can be captured and analyzed automatically?
(2) What propositions and observations about these systems can be made?
(3) Which methods for classifying and clustering of Web sites can be applied based on this automatically captured information?

Research phases

The research plan is broken down into three phases that are supported by the distinct components of the tool set (see Figure 1):
(1) Web mirroring;
(2) extracting the classification criteria; and
(3) analysis and clustering mechanisms.

These components for capturing and processing the empirical data are described in the following sub-sections. The proposed methodology comprises mirroring, script-based feature extraction, and five different analytical techniques. Each of the three phases has a clearly defined input and output interface to both the user and subsequent phases. Furthermore, the three phases can be run rather independently from each other. Therefore, future research may choose to employ a different implementation in one phase and keep the other implementations.

Web mirroring (A)
The collection of empirical data about the examined Web sites represents the first phase in this research. Since the World Wide Web is constantly undergoing evolutionary change (compare Bauer *et al.*, 1999; Bauer and Scharl, 1999), the computational and bandwidth requirements of a real-time analysis based on deployed systems would be very difficult to fulfil. The flexibility to modify existing and add

new classification criteria at a later point of time for the longitudinal studies is another argument for storing and archiving the original data.

The public availability of suitable Web mirroring tools allows substituting this functionality from the prototype tool with a corresponding interface implementation. After a short software evaluation, the tool HTTrack[13] appeared to provide the best implementation of the required functionality. Technical constraints and feasibility impose a number of restrictions to reduce the complexity of the mirroring task: only HTML files are downloaded (omitting memory-intensive multimedia formats such as image, sound or video files) and an upper limit of ten megabytes was imposed. Longitudinal studies require periodical data collection to reflect recent changes to the examined Web sites, thus mirroring and CD-ROM archiving of the following sectors are scheduled quarterly: information technology, retail banking, travel and tourism, environmental non-profit organizations, and academic institutions (see Figure 1).

Extracting the classification criteria (B)
The WebAnalyzer, implemented in Perl5, parses the HTML files of the mirrored Web sites, and computes the variables for the classification criteria listed in Table I. Variables for a particular Web site are assembled into a single vector. Thus a given subset of a certain sector can be represented as a two-dimensional matrix. Considering technical feasibility, the criteria and variables identified in Table I evolved from analyzing previous approaches, and from exploratory case studies. Almost all variables are measured on a numeric scale (with the only exception of the nominal frame attribute). The variables are categorized into three groups:
(1) content;
(2) interactivity; and
(3) navigation.

Additionally, the WebAnalyzer produces a single text file from the mirrored documents, which is stripped from the structural HTML markup, and stored in plain ASCII (American standard code for character interchange). This file serves as input for the exploratory textual analysis.

Figure 2 shows the homepage of 20/20 Vision, a grassroots activism organization based in the United States. Figure 3 presents part of the corresponding HTML code. The WebAnalyzer parses the HTML code and searches for markup tags to establish the values for the predefined parameters. The variables of total number of images and distinct number of images, for example, are derived from the occurrences and attributes of the <IMG>-tag (highlighted in Figure 3). Each occurrence of the tag increments the total number of images, while the distinct number of images is only increased when an image occurs for the first time – as indicated by the graphical resource's URL (uniform resource locator) in the SRC attribute of the <IMG> markup tag. Similarly, the variables about external and internal links are calculated by counting the <A> markup tags (also highlighted). A more detailed analysis of the URL in the HREF attribute distinguishes between relative, absolute, and outside links.

Analysis and clustering mechanisms (C)
The extracted raw information needs to undergo further analysis to associate semantic meaning. The classification of a group of Web sites into sub-sets requires a clustering mechanism. Depending on the employed method and the applicable input information, we distinguish five different analysis methods (see Figure 1):

Figure 2 Web information system of 20/20 Vision[14]

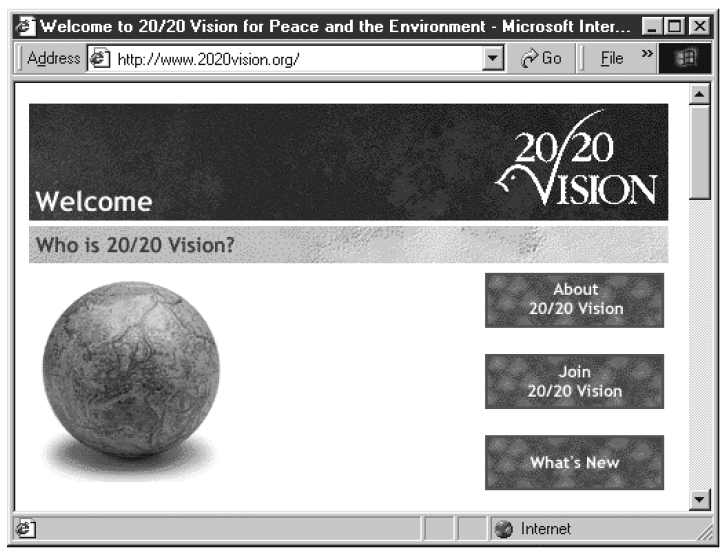


Figure 3 Excerpt of the HTML source code for Figure 2

```
<HTML>
<HEAD>
.
.
<TITLE>Welcome to 20/20 Vision for Peace and the Environment</TITLE>
</HEAD><BODY BGCOLOR="#ffffff">
<IMG SRC="pictures/welcomebanner.gif" WIDTH="464" HEIGHT="85"
ALIGN="BOTTOM" NATURALSIZEFLAG="3">
<BR><IMG SRC="pictures/whois.gif" WIDTH="464" HEIGHT="27"
ALIGN="BOTTOM" NATURALSIZEFLAG="3">
<BR>
<TABLE border=0 width=464><TR>
<TD align=left><IMG SRC="pictures/globe.gif">
<TD align=right><A href="about2020.html">
<IMG SRC="pictures/about2020.gif"></A><BR><BR>
<A href="https://secure.intr.net/2020vision/secure/join2020.html">
<IMG SRC="pictures/join2020.gif"></A><BR><BR>
<A href="whatsnew.html"> <IMG SRC="pictures/whatsnew.gif"></A>
</TD></TR></TABLE>
<BR>
<IMG SRC="pictures/grassdemoc.gif" WIDTH="464" HEIGHT="27"
ALIGN="BOTTOM" NATURALSIZEFLAG="3">
<BR>
.
.
</BODY>
</HTML>
```

(1) manual classification;
(2) textual analysis;
(3) statistical clustering;
(4) non-supervised neural network; and
(5) supervised neural network.

Any combination of the five suggested methods may be chosen as appropriate for different application domains. Each of these methods adds a particular perspective and specific quality attributes to the analysis. Combining and comparing the methods enables the validation of the models and derived conclusions.

Example: environmental non-profit organizations

The functionality and potential of the presented methods are illustrated in this section by applying them to the Web information systems of environmental non-profit organizations. The primary reasons for choosing this sector as an illustrative example are the (relatively) low complexity of Web sites, which are operated on a tighter budget compared to those of multinational corporations. Additionally, the limited number of sites in a smaller, clearly defined sector facilitates the interpretation of results. The names of the organizations and the corresponding URLs are listed in Table IV.

Manual classification

Manual classification tends to be more content-oriented than automated evaluations. The morphological chart in Figure 4 specifically was developed for Web information systems of environmental organizations, which is reflected in its six attributes. The morphological description specifies attributes of Web sites, together with their actual values. The attribute values are restricted to the relevant set provided in the morphological chart. The first attribute, strategy, is based on the organization's approach of how to change the world, either aggressively by putting pressure on industry or cooperatively by seeking mutual beneficial situations and joint projects. As far as the organization's goal is concerned, informative

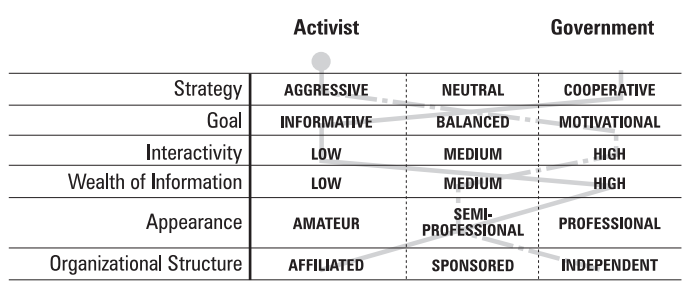
Table IV Sample of environmental Web sites

Name	URL
1. 20/20 Vision	http://www.2020vision.org/
2. Action Resource Center	http://www.arcweb.org/
3. Commission for Environmental Cooperation	http://www.cec.org/english/
4. Committee for a Constructive Tomorrow	http://www.cfact.org/
5. Conservation International	http://www.conservation.org/
6. Earth Communications Office	http://oneearth.org/fs_index.htm
7. Earth Island Institute	http://www.earthisland.org/
8. Earth Pledge Foundation	http://www.earthpledge.org/
9. Earth Summit Watch	http://www.earthsummitwatch.org/
10. Earthwatch Institute	http://gaia.earthwatch.org/
11. EcoNet	http://www.igc.org/igc/econet/
12. EnviroLink	http://www.envirolink.org/
13. Environmental Defense Fund	http://www.edf.org/
14. Environmental Law Institute	http://www.eli.org/
15. Friends of the Earth	http://www.foe.org/
16. Greenpeace International	http://www.greenpeace.org/
17. State of the World Forum	http://www.worldforum.org/
18. International Rivers Network	http://www.irn.org/
19. Natural Resources Defense Council	http://www.nrdc.org/nrdc/
20. Nature Conservancy	http://www.tnc.org/
21. One World	http://www.oneworld.org/
22. Rainforest Action Network	http://www.ran.org/intro.html
23. Sea Shepherd Conservation Society	http://www.seashepherd.org/
24. Sierra Club	http://www.sierraclub.org/
25. United Nations Environment Program	http://www.unep.org/
26. Wilderness Society	http://www.wilderness.org/
27. World Conservation Union	http://www.iucn.org/
28. World Resources Institute	http://www.wri.org/
29. World Watch Institute	http://www.worldwatch.org/
30. World Wide Fund for Nature	http://www.panda.org/

sites have to be distinguished from motivational ones, which are mainly concerned with activating the reader. Not surprisingly, interactivity plays a crucial role as well, especially for motivational sites. Independent of the organization's goal, the wealth of provided information and its (graphical) appearance represent two additional criteria to classify environmental Web information systems. Last, but not least, the structure of an organization – i.e. funded by a larger organization, by corporate sponsoring, or independently via direct membership programs – is directly reflected in the site's content. Indirectly, it also influences the other dimensions of the morphological chart.

The morphological approach allows identifying appropriate values for each attribute. In Figure 4, these values have been charted for activist versus government-oriented organizations (as indicated by the grayed lines). Naturally, not all the Web sites of a particular group will exactly follow this distribution. The identification of groups and their morphological attribute values, however, leads to better

Figure 4 Morphological chart for classifying environmental Web sites based on observable attributes



clustering and allows reverse allocation of organizations (by means of their Web information systems rather than organizational attributes).

Textual analysis

From a linguistic point of view, analyzing Web information systems is mainly concerned with syntax (relationship between words in sentences), semantics (meaning of the words), and pragmatics (associations between utterances and the circumstances of communicating), neglecting other established linguistic fields like phonetics, lexicology, or morphology. Automatic indexing and frequency computations ignore information of semantic or syntactic nature that usually is available to any reader, e.g. synonyms (two strings that have identical meaning), homonyms (one string that has more than one meaning) or the particular order of words within a text (Lebart, 1998). However, it is important to keep in mind that even this reduced set of textual data not only comprises the textual material itself but also includes abundant linguistic meta-data (words with rich semantic meaning are often allocated several pages in encyclopedic dictionaries).

The textual analysis is based on the ASCII output of the WebAnalyzer as described above. Since exploratory textual analysis is based on linguistic units like words or lemmas, additional computational tools are required (e.g. the commercially available tool WordSmith[15], which has been chosen by the authors). Cross-tabulations and lexical tables are used for representing the textual data retrieved from the World Wide Web. A number of transformations are needed prior to any statistical analysis. The analytical sequence includes the extraction of plain text from markup documents, the creation of the word list including lemmatization, and the application of a frequency threshold to limit the number of distinct words. On the basis of this word list, hierarchical cluster analysis and correspondence analysis are used to visualize the emergence and decay of topics of interest, both for longitudinal studies and for comparisons by sector.

Statistical clustering

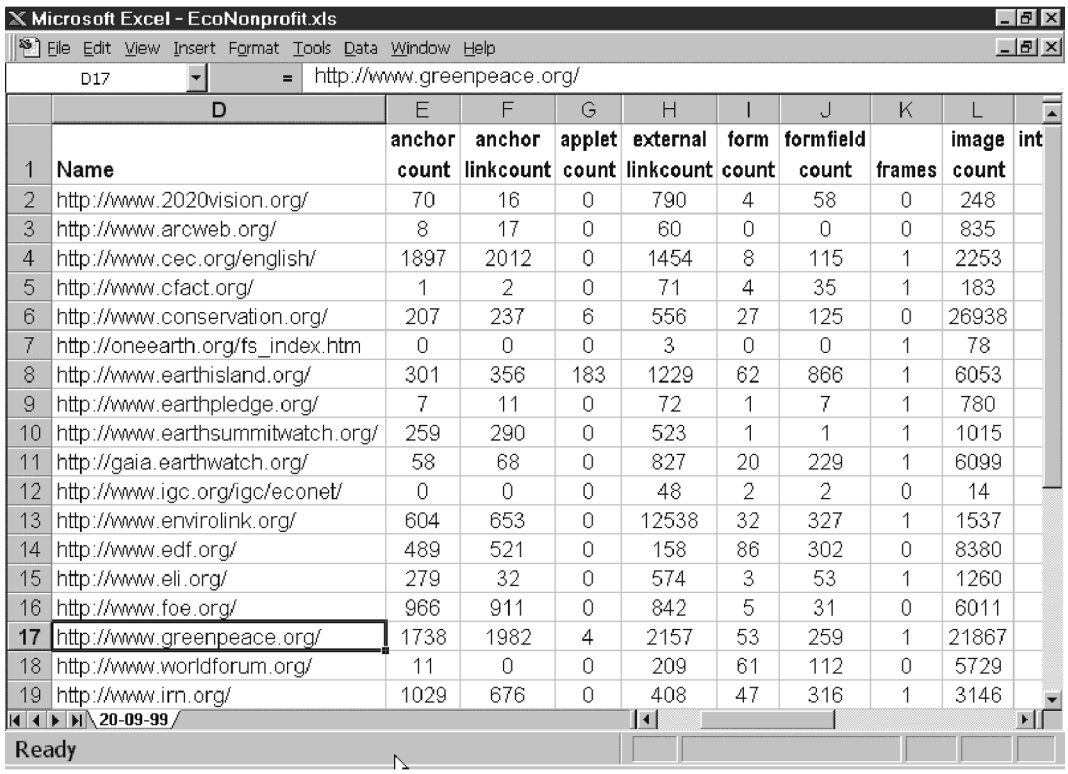
The aim of statistical clustering is to find groups of sites, where the members of each group are similar to the other members of the same group, and different from members of other groups. The most popular clustering algorithm is usually denoted as K-means (see, for example, Kendall *et al.*, 1983). This method requires the user to make several choices for parameters of the clustering algorithm, in particular the distance function, and the intended number of clusters.

The manual approach is based on the original Web sites, while the textual analysis utilizes the textual output of the WebAnalyzer. In contrast to both of these methods, statistical clustering and neural networks described in the following sections utilize the WebAnalyzer's quantitative output. Results are represented in a two-dimensional matrix as illustrated in Figure 5 (the numbers of rows and columns in the matrix table are reduced due to layout considerations; for a complete list of attributes see Tables I and IV). The matrix is a representation of the classification criteria vectors for each site. In the current implementation, the WebAnalyzer assembles the results in a comma-delimited text file (.csv), which includes the headers for each column (classification criteria), and a separate line for each Web site and its attribute values.

Non-supervised neural network

Neural networks represent a family of models for quantitative data analysis, rather than a single technique. They are composed of a large number of processing elements (= neurons) that are linked via weighted connections. Learning typically occurs by example through exposure to a set of correct input-output data, where the training algorithm iteratively adjusts the connection weights. These connection weights associatively store the knowledge necessary to solve specific problems. Neural networks as universal approximators work very well for discovering regularities within a complex set of patterns, where the volume of the data is very great, the non-linear relationships between variables are only vaguely understood, or the relationships are difficult to describe with conventional approaches. Thus they provide a promising analytical alternative to statistical techniques, which are often limited by strict assumptions of normality, linearity, or variable independence.

Figure 5 Top left corner of the WebAnalyzer's two-dimensional output matrix



	D	E	F	G	H	I	J	K	L	
1	Name	anchor count	anchor linkcount	applet count	external linkcount	form count	formfield count	frames	image count	int
2	http://www.2020vision.org/	70	16	0	790	4	58	0	248	
3	http://www.arcweb.org/	8	17	0	60	0	0	0	835	
4	http://www.cec.org/english/	1897	2012	0	1454	8	115	1	2253	
5	http://www.cfact.org/	1	2	0	71	4	35	1	183	
6	http://www.conservation.org/	207	237	6	556	27	125	0	26938	
7	http://oneearth.org/fs_index.htm	0	0	0	3	0	0	1	78	
8	http://www.earthisland.org/	301	356	183	1229	62	866	1	6053	
9	http://www.earthpledge.org/	7	11	0	72	1	7	1	780	
10	http://www.earthsummitwatch.org/	259	290	0	523	1	1	1	1015	
11	http://gaia.earthwatch.org/	58	68	0	827	20	229	1	6099	
12	http://www.igc.org/igc/econet/	0	0	0	48	2	2	0	14	
13	http://www.envirolink.org/	604	653	0	12538	32	327	1	1537	
14	http://www.edf.org/	489	521	0	158	86	302	0	8380	
15	http://www.eli.org/	279	32	0	574	3	53	1	1260	
16	http://www.foe.org/	966	911	0	842	5	31	0	6011	
17	http://www.greenpeace.org/	1738	1982	4	2157	53	259	1	21867	
18	http://www.worldforum.org/	11	0	0	209	61	112	0	5729	
19	http://www.irn.org/	1029	676	0	408	47	316	1	3146	

If the task of neural learning is to discover regularities without specified output patterns or supervising structures, the corresponding learning process is called non-supervised or self-organizing learning (Inform, 1997; Masters, 1993). The most famous algorithm geared toward non-supervised learning is the Kohonen network, named after its inventor, Teuvo Kohonen. Following early approaches, which successfully identified phonemes from continuous speech (Kohonen, 1988, 1997), the Kohonen self-organizing map has been used for diverse purposes. The Kohonen algorithm effectively projects the original data onto a subspace of (usually) two dimensions, such that similar data are placed close to one another. This representation is referred to as a topographical map, which denotes a two-dimensional array of neurons fully connected with the input vector, but without lateral connections (Silipo, 1999).

Supervised neural network

Supervised neural networks are typically organized in layers, each of them comprising a number of interconnected nodes. Learning rules modify the weights of the connections according to the input patterns that they are presented with. Once clusters have been established (either via statistical algorithms, or via non-supervised neural learning), it becomes feasible to develop neural networks using a supervised learning algorithm such as backpropagation. This algorithm, independently reinvented multiple times, performs a gradient descent within the solution's vector space along the steepest vector of the error surface (Silipo, 1999). Once a neural network is trained to a satisfactory level, it may be used as an analytical tool on other data (Pyle, 1999; Mena, 1999).

Although self-organizing maps can be used for classification into previously established clusters as well, they are not optimal for that application. A properly trained feed-forward network would be expected to be faster and more accurate. Applying hybrid approaches, which combine the adaptive features of neural networks with the modeling flexibility of fuzzy logic, has become increasingly popular in various disciplines (e.g. Scharl, 1999). Especially for non-numeric classification

criteria this seems to be a promising alternative, and would eventually allow the integration of manual and automated classification methods. Therefore, the authors are currently in the process of testing the domain-specific suitability of FuzzyTech 5.01, a neurofuzzy-shell that employs a supervised neural component to specify the rules and membership functions of the fuzzy inference engine (Inform, 1997)[16].

Subsequently, the assessment tool will use an adaptive neurofuzzy architecture to determine the appropriate category (output vector) for any given Web site. The optimization of the neural component integrated into the 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 (non-supervised 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 Web site evaluation are concerned.

Conclusions and further research

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 sectors. 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 system. While fuzzy modeling maintains the transparency of the system's internal calculations, the neural component ensures maximum flexibility and continued optimization, eliminating the need for predefined and often questionable mathematical relationships.

Notes

1 http://www.commerce.net/news/press/19981026.html
2 http://www.surveysite.com/
3 http://www.w3.org/PICS/
4 http://www.yahoo.com/
5 http://www.excite.com/
6 http://www.unc.edu/~elliott/evaluate.html
7 http://validator.w3.org/
8 http://www.websitegarage.com/
9 http://www.fritz-service.com/
10 http://www.cast.org/bobby/
11 http://www.spec.org/
12 http://www.tpc.org/
13 http://e3000.ensicaen.ismra.fr/~roche/httrack.html
14 http://www.2020vision.org/
15 http://www.liv.ac.uk/~ms2928/index.htm
16 http://www.fuzzytech.com/

References

Bauer, C. (1998), *Internet und WWW für Banken: Inhalte, Infrastrukturen und Erfolgsstrategien*, Gabler, Wiesbaden.

Bauer, C. and Scharl, A. (1999), "Advanced design of Web information systems based on dominant and emerging Web communication patterns", Paper presented at the The Fifth Australian World Wide Web Conference (AusWeb'99), Ballina, Australia.

Bauer, C., Glasson, B. and Scharl, A. (1999), "Evolution of Web information systems: exploring the methodological shift in the context of dynamic business ecosystems", in Romm, C. and Sudweeks, F. (Eds), *Doing Business On The Internet: Opportunities and Pitfalls*, Springer, London, pp. 35-52.

Booz-Allen and Hamilton (1999), *Corporate Internet Banking: A Global Study of Potential*, available at: http://www.bah.com/press/net_corpbank.html (accessed 22 May 1999), Booz-Allen and Hamilton.

Bucy, E.P., Lang, A., Potter, R.F. and Grabe, M.E. (1999), "Formal features of cyberspace: relationships between Web page complexity and site traffic", Paper presented at the 49th Annual Conference of the International Communication Association (ICA-99), San Francisco, USA.

Galliers, R.D. (1993), "Research issues in information systems", *Journal of Information Technology*, Vol. 8 No. 2, pp. 92-8.

Hansen, H.R. (1996), *Klare Sicht am Info-Highway: Geschäfte via Internet & Co.*, Orac, Vienna.

Hansen, H.R. and Tesar, M.F. (1996), "Die Integration von Masseninformationssystemen in die betriebliche Informationsverarbeitung", Paper presented at the Fachtagung "Data Warehouse" an der Gerhard-Mercator-Universität GH Duisburg, Duisburg, Germany.

Hoffman, D.L., Novak, T.P. and Chatterjee, P. (1997), "Commercial scenarios for the Web: opportunities and challenges", *Journal of Computer-Mediated*

Communication, Vol. 1 No. 3, available at: http://www.ascusc.org/jcmc/vol1/issue3/

Inform (1997), *FuzzyTech 5.0 User's Manual*, Inform Software Corporation, Aachen.

Jutla, D., Bodorik, P., Hajnal, C. and Davis, C. (1999), "Making business sense of electronic commerce", *Computer*, Vol. 32 No. 3, pp. 67-75.

Kendall M., Stuart, A. and Ord, J.K. (1983), *The Advanced Theory of Statistics. Volume 3: Design and Analysis, and Time Series*, Charles Griffin, London.

Kohonen, T. (1988), "The 'neural' phonetic typewriter", *Computer*, Vol. 21 No. 30, pp. 11-22.

Kohonen, T. (1997), *Self-organizing Maps*, 2nd ed., Springer, Berlin.

Lebart, L. *et al.* (1998), *Exploring Textual Data* Kluwer Academic Press, Dordrecht.

Mahler, A. and Göbel, G. (1996), "Internetbanking: Das Leistungsspektrum", *Die Bank*, Vol. 8, pp. 488-92.

Masters, T. (1993), *Practical Neural Network Recipes in C++*, Academic Press, San Diego.

McMillan, S.J. (1999), "The microscope and the moving target: the challenge of applying a stable research technique to a dynamic communication environment", Paper presented at the 49th Annual Conference of the International Communication Association (ICA-99), San Francisco, USA.

Mena, J. (1999), *Data Mining Your Website*, Digital Press, Boston, MA.

Nunamaker, J.F., Chen, M. and Purdin, T. (1990-91), "Systems development in information systems research", *Journal of Management Information Systems*, Vol. 7 No. 3, Winter, pp. 89-106.

Olsina, L., Godoy, D., Lafuente, G.J. and Rossi, G. (1999), "Specifying quality characteristics and attributes for Websites", Paper presented at the First ICSE Workshop on Web Engineering (WebE-99), Los Angeles, USA.

Psoinos, A. and Smithson, S. (1999), *The 1999 Worldwide Web 100 Survey*, London School of Economics, London.

Pyle, D. (1999), *Data Preparation for Data Mining*, Morgan Kaufmann, San Francisco, CA.

Scharl, A. (1997), *Referenzmodellierung kommerzieller Masseninformationssysteme: Idealtypische Gestaltung von Informationsangeboten im World Wide Web am Beispiel der Branche Informationstechnik*, Peter Lang, Frankfurt and Vienna.

Scharl, A. (1999), *Neurofuzzy-Hybridsysteme: Theoretische Grundlagen, Vergleich mit statistischen Problemlösungsklassen und Anwendung im Rahmen leistungsdiagnostischer Fragestellungen*, WUV-Universitätsverlag, Vienna.

Scharl, A. and Brandtweiner, R. (1998), "A conceptual research framework for analyzing the evolution of electronic markets", *International Journal of Electronic Markets*, Vol. 8 No. 2, pp. 39-42.

Selz, D. and Schubert, P. (1997), "Web assessment: a model for the evaluation and assesment of successful electronic commerce applications", *International Journal of Electronic Markets*, Vol. 7 No. 3, pp. 46-8.

Silipo, R. (1999), "Neural networks", in Berthold, M. and Hand, D.J. (Eds), *Intelligent Data Analysis: An Introduction*, Springer, Berlin, pp. 217-68.

Tesch, R. (1990), *Qualitative Research: Analysis Types and Software Tools*, Falmer Press, New York, NY.

Witherspoon, E.M. (1999), "A pound of cure: a content analysis of health information on Web sites of top-ranked HMOs", Paper presented at the 49th Annual Conference of the International Communication Association (ICA-99), San Francisco, USA.

Further reading

Schubert, P. and Selz, D. (1999), "Web assessment – measuring the effectiveness of electronic commerce sites going beyond traditional marketing paradigms", Proceedings of the 2nd Hawaii International Conference on Systems Sciences, Hawaii, USA.

Shneiderman, B. (1997), "Designing information-abundant Web sites: issues and recommendations", *International Journal of Human-Computer Studies*, Vol. 47 No. 1, pp. 5-29.