7 Mining Web Collections

Andreas Aschenbrenner[1] and Andreas Rauber[2]

[1] Goettingen State and University Library
[2] Vienna University of Technology

7.1 Introduction

First contact with Web archiving is often with the technological issues in collecting Web material, which is discussed in the first part of this book. However, it is also one of the key messages of this book that Web archiving and the stewardship of Web material over the long term entails an array of tasks and functions. Thus, after the first part of this book on the more technical issues in building Web archives, this part discusses the usage of Web archives followed by their preservation in the next part. It is the ambition of this chapter to highlight the intricate interrelations between Web archive construction, usage, and preservation, to illustrate the myriad of issues involved in Web archive usage, and to convey the importance of planning and organisation of Web archives with respect to their later usage.

Usage of Web archives intuitively implies a module building on an existing Web archive that allows access in a similar manner to the way we access the current Web with the additional dimension of the time. Thorsteinn Hallgrímsson describes such a module in Chap. 6, and the popular access point to the Internet Archive, the Wayback Machine is highlighted in a respective case study in the last part of this book. Besides such access modules, however, there are a wealth of other tools and techniques for web archive usage, supporting access to web archives and the analysis of their content. Additional means for access and analysis enhance the value of a Web archive enormously, and may even attract entirely new target audiences. These means are to a great extent subsumed in the broad field "Web mining". What are the (hidden) sources in a Web archive and what usage scenarios can we infer from them? This chapter describes what "Web mining" refers to and outlines various usage scenarios based on selected case studies that excel in unlocking the hidden information of Web archives.

The creation of novel services and the attraction of more clients create enhanced business value of the Web archives, yet they also entail additional tasks and responsibilities. Usage is more than a mere add-on module on any Web archive, and demands adequate planning and effort. Particular usage schemes may demand collecting specific Web material and may raise requirements for their management and preservation. Similarly, the needs of the archive's user group need to be taken into account from the outset. This shows the tight integration of a Web archive's usage in all its other functions including collection and preservation as well. By exploring possible usage schemes and specifying the Web material they are based on, this chapter therefore aims to support delineating a Web archive's dedicated community and its scope, defining requirements for collection and archival management, as well as for the preservation of the collections.

Moreover, exploring possible usage scenarios and selecting the appropriate ones is the basis for establishing the envisioned role of the Web archive within an organisation or for society. This, in turn, is a prerequisite for shaping the organisational structure of the archive, eliciting funding sources, and establishing requirements for the technical environment. All these points will help to ensure the archives' sustainability in the various dimensions of the word: organisational, financial, as well as technological sustainability. For all these reasons the exploration of usage mechanisms is at the core of any Web archiving initiative.

In addition to the challenges in constructing a Web archive and maintaining a successful service, a Web archive and specifically the combination of different data sources and mining techniques may have a significant social impact that needs to be considered from the outset. This regards most notably privacy concerns relating to Web archives equally as to any data collection initiative. Leaving this issue to (national) legislation falls short of resolving this ethical issue; it demands an open discussion involving all stake holders, including those responsible for the Web archive, data creators and providers, as well as the users. Without any specific and general recommendation being available at this point of time, Web archivers and Web miners should keep in mind the social repercussions of the analyses they are performing, and incorporate strategies to calm privacy concerns by anonymisation, time-delayed access, and by trying to foresee the potential consequences of the kind and amounts of data they amass – while at the same time obviously trying to keep their archive and analyses comprehensive and powerful to live up to both its scientific and social value.

Web mining is a broad field that embraces a variety of techniques for the extraction of patterns from and for the analysis of a web archive's holdings. It is a very dynamic scientific field that grows with the proliferation of information and services on the Internet, constantly adapts to new Internet technologies, and incorporates new analysis techniques. Indeed, various attempts to structure the field according to the Web material used or the mining technique were dwarfed by the field's rapid development. An early account (Cooley et al. 1997), for example, distinguishes between Web content mining and Web usage mining. This classification differentiates along the different kinds of material used; the actual Web pages on the one hand and access log files on the other. Only in the late 1990s techniques such as the PageRank Algorithm (Brin and Page 1998) employed by the search engine Google[1] were developed, which draw on the links embedded in Web content rather than the content data itself. This development added another area of Web mining, namely Web structure mining, which later descriptions of the field accounted for. Today, techniques to analyse Internet infrastructure data and Web server transfer meta-data enhance the list of Web material apt for Web mining, and the emerging field of network analysis is a new addition to Web mining techniques. While this chapter has no ambition to develop an updated typology for Web mining, it will present an updated view on the issues in Web mining with its requirements and possibilities. This account focuses purely on Web mining use cases that are currently applied or could potentially be applied to Web archives while omitting techniques such as wrapper induction (Kushmerick 2000) or other, purely Information-Retrieval oriented methods and projects. While it is impossible to provide a full account now or in the future, this chapter provides a broad synopsis of the field and equips the reader with the necessary triggers for shaping her or his Web archiving initiative. The upcoming section provides a general overview of the various kinds of Web material that are employed by Web mining techniques either isolated from each other or in any combination. These diverse resources may be considered for inclusion in Web archives to allow for existing Web mining applications to be installed and foster research. A subsequent description of selected case studies illustrates the range of conceivable usage scenarios and potential services.

7.2 Material for Web Archives

Various kinds of data offer themselves for inclusion in a Web archive. Some of these data are very obvious; some of them less so, but they can

[1] Google. www.google.com

easily be collected alongside other activities; and some data may be available for some organisations, however, more difficult to obtain for others. The following list gives an idea of the diversity of Web material at our finger tips. These various data are employed by the use cases compiled in the "Use Cases" section later.

7.2.1 Web Pages

The first and most important material for Web archives obviously are the actual Web pages. Their collection has been exhaustively covered in this book's Chaps. 1, 3–5. Stressing again the myriad of data formats available on the world wide Web, the pioneering Swedish Web archiving initiative Kulturarw3 (Arvidson et al. 2000) found 465 different file formats in their 10th sweep of the Swedish Web space, which took almost ten month from August 2002.[2] The five most frequent file types, which covered 96% of the files captured in this sweep, are HTML (HyperText Markup Language) documents, images in GIF and JPEG format, as well as plain text and PDF documents. Other types include various sorts of multimedia and presentations.

The most characteristic file type for the world wide Web and, indeed, the most prevalent type that constitutes almost half of all the online available documents, are documents written in HTML that are interpreted by Web browser applications. Besides static Web pages, HTML documents may be dynamically generated by a Web server to satisfy a specific user request. Web pages with a file extension PHP, CGI, or the like point to dynamically generated pages. All these documents may be marked up in HTML versions up to HTML 4 or the eXtensible HyperText Markup Language XHTML.[3] HTML is the document type in the focus of most current Web mining activities. In the future, Web mining may incorporate techniques from multimedia mining (Zaiane 1999) and other data mining areas, but at this point of time the field is mainly focusing on the distinctive properties of HTML.

When dissecting the name "HyperText Markup Language" we find its first main feature to be hypertext. Hypertext has links embedded in its textual content that enable the reader to control the flow of reading by jumping from one page to another in a self-chosen order and fit to her needs.

[2] Kulturarw3, Long time preservation of electronic documents. The web archiving initiative of the Swedish Royal Library. http://www.kb.se/kw3/

[3] The World Wide Web Consortium (W3C) HyperText Markup Language (HTML). http://www.w3.org/MarkUp/

This non-linear way of reading has entirely changed traditional publishing paradigms. The hyperlink in one page pointing to another page also marks a relation between the two pages. Based on this relation a graph of a website or, theoretically, the whole Web can be drawn. In such a Web graph a page is a vertex, and each link within the page is an edge to another vertex. Viewing the Web as a graph offers a wholly new perspective that largely fades out the textual content of a page and focuses on the relations between Web documents.

HTML's second main feature is its semi-structured nature, as the term "markup" indicates. Thereby, each HTML document consists of a number of chunks that are started and ended by a, respective, tag. For example the highest level heading of a page is started with <H1> and terminated with </H1>. Tags may be recursively embedded within each other. Structuring the HTML document in clearly defined chunks allows automatic tools a basic level of machine comprehension. It is HTML's semi-structured nature that greatly influences and enhances the possibilities of Web mining applications.

7.2.2 Metadata

Metadata – chiefly defined as data about data – also play a pivotal role on the Web and consequently also in Web archives. These background data conveying a wealth of information in addition to the obvious content of the Web objects can be gleaned from the Web or they are produced through the application of mining techniques.

There are two main metadata types of interest with regard to Web mining: metadata about the object from the Web, as well as technical metadata in context with the transmission of the object via the Internet. Metadata about a Web object can be obtained from various sources. One major source for metadata is, of course, the object itself including its date of creation and object size. The file format can be extracted, either by simply looking at the file extension of the object or by applying tools like JHOVE[4] that identifiy the format based on internal properties of the file. The human language used in textual documents is another metadata item to be gained from the object itself by help of specialised external language detection algorithms, besides a number of other conceivable metadata items. As a further metadata source, the object may carry some limited metadata in dedicated

[4] JHOVE – JSTOR/Harvard Object Validation Environment. http://hul.harvard.edu/jhove/jhove.html

fields, such as for example Dublin Core[5] mark-up. Tools such as UKOLN's DCdot[6] can extract these metadata easily. Due to the lack of a central control in the Web, however, there is rarely consistent metadata in objects, if at all. This source will hence be only fruitful for organisations that have the authority to centrally impose a specific structure on their Web material including specific metadata. This source is only of limited use for initiatives that tackle the open Web or corporate Web mining initiatives unable to control the addition of metadata to their Web material. Lastly, communication with the Web server offers some limited but important object metadata including its location (i.e., the URL) and the MIME type of the object.

Communication with Web servers also yields technical metadata in context with the transmission of the object via the Internet. When an object is requested by a Web browser, the browser typically opens a HTTP connection to port 80 of a Web server and requests a specific document. The response from the Web server contains the requested object if available, as well as additional metadata including information about Web server software (e.g., "Apache/1.3.26 (Unix) PHP/4.3.3 mod_ssl/2.8.9 OpenSSL/-0.9.6c"), the connection status in a standardised code, and the MIME type of the transferred Web object. As we will see in the use cases later, this information gives some valuable information about the technology infrastructure in the Internet.

7.2.3 Usage Data

The primary source for usage data are server logs (Mobasher 2004). Every time a user sends a request to a Web server, the Web server protocols that together with some additional data about the request and where it came from. A standard server log format is the W3C Extended Log File Format (Hallam-Baker and Behlendorf 1996); others such as Microsoft IIS Log Files hold similar information with some minor format variations. A log file may include along with transmission data such as date of the request and number of bytes transmitted also user information such as the user IP address, user name or the computer name of the user. Web server logs are easy to get for corporate Web archiving initiatives. External Web archiving initiatives will, however, only get this data with the cooperation of the respective organisation and the respective Web server administrator.

[5] Dublin Core. www.dublincore.org

[6] DCdot – UKOLN (UK Office for Library Networking) Dublin Core metadata editor. http://www.ukoln.ac.uk/metadata/dcdot/

Other usage data may include personal user information created by a specific Web service. For example, some websites may be restricted to registered users and allow the user to create a personal user profile. Online services may record a history of past user navigation patterns, and perhaps include explicit user ratings on products as a basis of collaborative filtering systems. The Web navigation service Alexa,[7] for example, builds on user ratings and traffic rankings. These data are, however, reserved to the respective Web site hosts providing such a special service. Passing this data on to third parties may raise severe data protection concerns.

7.2.4 Infrastructure Data

As mentioned earlier, the technical metadata retrieved in the course of HTTP transmissions offers some limited information about the technology setup of Web servers. This data, however, still does not reveal information about the overall infrastructure of the Internet and how the various Web servers are interconnected. Data that reflects the Internet infrastructure are the routing tables. When data is routed through the Internet it is passed from one local network – also referred to as an autonomous system – to the neighbouring unit until it reaches the destination. Routers are the entities that forwards the data packages, and routing tables tell them which neighbouring autonomous system is next in the data package's path to its final destination. The Border Gateway Protocol (BGP) (Rekhter and Li 1995) is a protocol for exchanging routing information between routers. BGP data contains some resources that reflect Internet connectivity and the overall technology infrastructure at the Internet backbone connections, including IP addresses, autonomous system numbers, CPU cycles in routers and bandwidth consumed by routing update traffic (Brody and Hickman 2000). BGP data is, however, difficult to obtain. There are several initiatives that archive and provide BGP related traffic data. For example, the National Laboratory for Applied Network Research's Measurement and Operations Analysis Team (MOAT)[8] provides BGP data from almost 50 contributors who deposited BGP tables between November 1997 and March 2001. Other such initiatives include the Internet Traffic Archive,[9] as

[7] Alexa. http://www.alexa.com/

[8] National Laboratory for Applied Network Research (NLANR), Measurement and Network Analysis Group: Network Analysis Infrastructure (NAI). http://moat.nlanr.net/infrastructure.html

[9] ACM SIGCOMM: Internet Traffic Archive. http://ita.ee.lbl.gov/index.html

well as the Ripe NCC Routing Information Service[10] that provides a myriad of historical information about Internet traffic.

However, while BGP captures a detailed view of the centre of the Internet, the connectivity of the periphery is rarely captured. Active probing of Internet connectivity by using the Traceroute tool may yield a broader coverage, even though it fails to provide data as detailed as the BGP routing tables. Traceroute is a network administration tool capable of determining the route data packets take to reach a destination host. Tools based on the Traceroute utility can be implemented by anybody (though uncoordinated massive scanning from numerous initiatives would probably hamper global data traffic and annoy network administrators). A comprehensive archive has already been set up by the Internet Mapping Project (Cheswick and Burch, 2004) that started to collect Traceroute traces in 1998 and has the long-term mission of acquiring and saving Internet topological data.

7.3 Other Types of Information

This list of Internet data types apt for Web archiving is, of course, not exhaustive. Over time a myriad of other Web material is likely to emerge only to perish again after being superseded by the next generation data type. Already now it is hardly possible to keep track of the host of different data types on the Internet. Applications and protocols such as the IRC chat channel, Multi User Dungeons and other online games, newsgroups, RSS (Really Simple Syndication Web content and distribution), mailing lists, and many others more are excluded from the above succinct and general list for the impossibility to provide a full account. Material considered for inclusion into a Web archive may even cover extraneous sources, such as monitoring (e.g., filming) user sessions within specialized user study labs, to capture the interactive nature of some Web applications, online games and other highly interactive elements on the Web.

Also, the accessibility of specific data may vary between Web archiving initiatives. Each Web archive therefore needs to identify for itself the data types, the scope of collection and quality requirements to fulfil its mission and to best serve its dedicated community. Some archives may choose to specialise on a single, perhaps one of the more rare data types. The Netscan Project,[11] which maps the social geography of usenet news, is exemplary for the valuable service and exciting research such a specialised archive

[10] Réseaux IP Européens (RIPE): Routing Information Service (RIS).

[11] Netscan Project. http://netscan.research.microsoft.com/

may provide. Other initiatives may choose to collect a variety of different data types, for a variety of different services or perhaps in anticipation of possibilities for future usage. In the latter case, the regular environmental scan by the National Library of Australia (Philips 2003) may be a model for other initiatives struggling to keep their corporate selection and collection policy updated in the rapidly changing environment.

As the use cases in the next section describe, Web mining techniques mostly need a very particular data chunk from the various Web material described earlier. Some techniques combine a variety of different Web material, yet they focus on specific data elements such as the name of the Web host in an external Web link. Web mining usually involves a preprocessing stage, in which the original data is tailored to the specific needs of the mining technique. In this stage specific data may be converted or augmented with other, external data, and other data may be dumped as it is not useful in the particular case. While these processed data are all needed for a specific data mining technique, the original material needs to be archived as well. Without the original data an initiative is confined to a small set of mining techniques. Moreover, perhaps a new version of a mining technique demands additional data elements. A Web mining initiative should therefore avoid obstructing the evolution of mining techniques and the development of the archive, and rather preserve the original data.

In addition to the original data, the context of the data to be preserved in a Web archive needs to be captured in many cases as well. Without exhaustive documentation future generations will hardly be able to make sense of the material. For infrastructure data, for example, documentation of the local host configuration is needed, and also of the surrounding Web environment (e.g., prevalent applications and prevailing protocols at that time). As another example, in the case of usage data real-world promotional activities may impact hugely on the user access rate and consequently the Web logs, so even a record of an organisation’s real-world activities could support future interpretation of the material collected by a Web archiving initiative. The preservation of context data alongside the original data could, of course, escalate easily and become unfeasible, yet it should be considered and clearly defined by the Web archiving initiative when establishing its collection policy.

7.4 Use Cases

The different kinds of data listed in the last chapter form the basis for a variety of applications. Some initiatives building on such data from Web archives are presented in the following. They feature both initiatives that focus solely on one type of data and others utilising a combination of them. A combination of different data types allows Web archives to serve a larger range of user communities and facilitates the application of entirely new, perhaps more reliable, mining techniques.

The following use cases are selected to convey the scale of inherently different mining techniques that may be built on Web archives. They highlight the valuable services a Web archive may fulfil. Indeed, Web mining techniques can unveil valuable information pertaining to all walks of life. Future Web archives may provide indispensable services for research, business and private life in all conceivable areas from research in sociology and mathematical theory to marketing analysis and public opinion tickers.

7.4.1 Analysing Web Content and Web Technology

With respect to the usage of Web archives, several projects are developing access interfaces that allow users to search and surf within such an archive, such as, e.g., the Internet Archive's Wayback Machine, or the access tools for the Nordic Web Archive (NWA). These tools provide for each URL a timeline listing the dates when this specific URL was added to the archive, i.e., which versions of the respective file are available.

Going beyond the mere navigation within the archive as a mirror of the World Wide Web existing at the respective times, several projects take a more structured approach to storing and analysing the Web. The Web Archaeology project (Leung et al. 2001) studies the content of the World Wide Web using a variety of content representations, referred to as features, including *links* capturing connectivity, *shingleprints* capturing syntactic similarities, and *term vectors* capturing semantic similarities. The Mercator Extensible Web Crawler is used for large-scale data acquisition, and specific database models were developed at the second layer of the system architecture for storing the feature databases. Various tools are added to the top layer of the system architecture to facilitate specific types of analysis, such as, e.g., in the Geodesy project trying to discover and measure the structure of the Web.

Another Web page repository is being built within the WebBase project at Stanford University, addressing issues such as the functional design, storage management, as well as indexing modules for Web repositories (Hirai 2000). The main goal of this project is to acquire and store locally a subset of a given Web space in order to facilitate the performance execution of several types of analyses and queries, such as page ranking, and information retrieval. However, it limits its scope to the archiving of one copy of each page at a time, thus providing no historisation, and focuses on HTML pages only.

When it comes to analysing large amounts of data in a flexible manner, data warehouses (DWH) have evolved into the core components of decision support systems (Kimball, 2002). A DWH is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of decision-making processes (Inmon 1992). Rather than storing data with respect to a specific application, the information is processed for analytical purposes, allowing it to be viewed from different perspectives in an interactive manner. It furthermore integrates information from a variety of sources, thus enriching the data and broadening the context and value of the information.

The primary concept of a DWH is the separation of information into two main categories, referred to as facts and dimensions, respectively. Facts are the information that is to be analysed, with respect to its dimensions, which often reflect business perspectives, such as a geographic location, evolution over time, product groups, merchandising campaigns, or stock maintenance. The DWH may be envisioned as a multi-dimensional data cube. Using on-line analytical processing (OLAP) tools, this data cube allows to interactively drill-down, roll-up, slice and dice, view and analyse the data from different perspectives, and to derive ratios and compute measures across many dimensions. These OLAP operations assist in interactive and fast retrieval of 2D and 3D cross-tables and chart-table data from the cube, which allow convenient querying and analysis of a Web data storage.

This technology is used to analyse web collection data in the WHOWEDA project, pursued by the Web Warehousing and Data Mining Group at the Nanyang Technological University in Singapore (Bhowmick et al. 2000). Within this project a DWH stores consecutive versions of Web pages, adding a time dimension to the analysis of content and link structure. URL, size, date of last modification (and validity, with respect to subsequent visits to a given site), size, etc. are stored together with the content and structure of a document. Furthermore, link information, as well as the position of links within documents are recorded and made available for further analysis. Although a more structured approach to the analysis of Web pages is taken within the scope of this project, it primarily focuses on a detailed analysis and representation of the content of the documents.

Technologically similar, the Austrian on-line archive processing project (AOLAP) (Rauber et al. 2002) uses DWH technology to analyse the data acquired by the Austrian on-line archive (AOLA) (Rauber and Aschenbrenner 2001). The archive consists of Web pages, including all types of files as collected by the harvesting software, and rests on tape archives organised primarily according to domain names. In addition to the actual pages, meta-information that is provided or created during the crawling process, is collected and stored as part of the archived files. This includes information provided as part of the HTTP protocol as well as other information provided by the server, such as the server software type and version, the operating system used by the server, date and time settings at the server, as well as last-modified dates for the respective file being downloaded.

The information extracted from the pages includes *file types* based on file extensions, *file size*, internal and external *links*, information about *frames*, *e-mail addresses* and interactive *forms* used in the case of HTML files, *date of last modification*, and others. With respect to the various domains it mainly concentrates on *IP addresses* and thus *network types*, *operating system* and *Web server software* information. Furthermore, AOLAP integrates information from other sources, to enrich the data provided by the harvesting system. Specifically, a set of WHOIS servers is used to provide geographic location information of Web service registrars, alias names, etc. The information is further transformed and loaded into a relational DBMS using a star-model like design for the data storage. Based on this model a multi-dimensional cube is created which can further be used for interactive analysis, such as the distribution of file types across different Web servers or link structure analysis.

The number of file types encountered in the Web archive is highly relevant with respect to the preservation of the archive that is, keeping the pages viewable in the near and far future. The number of types also represents a good mirror of the diversity of the Web with respect to the technologies employed for conveying information. Overall, AOLAP encountered more than 200,000 different types of files based on their extensions, and more than 200 different types of information representation when using the MIME type as the indicative criterion. However, the quality of the information provided this way is rather low, as a large number of both file extensions as well as MIME types are actually invalid, such as files with extensions .htmo, .chtml or .median, .documentation. While the majority of file extensions encountered definitely are erroneous, these are indicators of serious problems with respect to preserving that kind of information, as well as the need to define solutions for cleaning this dimension to obtain correct content type descriptors.

Analysing the link structure confirms an intuitive tendency, namely a high inter-linkage within each respective domain, i.e., .com sites linking mostly to other .com sites, .cc linking within .cc and so on. However, there are also some interesting exceptions to this rule, which can be interactively analysed by drilling down in the respective domains.

Since AOLA collects a huge range of metainformation during the crawl, these various types of information can be put in relation to each other, such as for example the distribution of different Web server types across the various domains, and the evolution of market shares of Web servers.

Obviously, further types of information can be extracted from the Web pages and integrated into a DWH (e.g., automatic language detection methods) covering in larger detail additional technological information, such as the usage of cookies, embedded Java applets, Flash plug-ins, encryption, and others. Furthermore, being able to analyse the content-based dimension of a Web archive provides the basis for subject gateways on a variety of topics and with a historic dimension.

Due to the flexibility offered by the DWH-based approach, these systems can be used in a wide range of Web archive utilisation scenarios, both for archive maintenance, as well as for exploiting the information constituted by the archive. The information obtainable via these DWH-based approaches is similar to statistics computed by some of the longer-running projects in this field, such as the Swedish Kulturarw3 project, which has available data from several complete runs since 1996. These statistics show, for example, the first traces of XML documents in early 1999 and reveal how XML documents have been increasing in number and as a share of Web documents available from 1999 to date. Another fascinating example is the surprisingly sudden victory of PDF over the Postscript file-format within about a year in 1998.

The main benefit of the proposed DWH approach for Web archive analysis lies in the flexibility with which interactive analysis of the archive can be performed. Contrary to most current approaches, the focus of this type of analysis is not primarily on the content of the data, but rather on meta-information about the data, as well as data about the technologies used to provide a given service.

7.4.2 Exploring Web Communities

Within a minimal time-span the Web has become the communication and publishing space of modern society. It grew from an experimentation space for scientists and computer geeks in its early times, to the all-inclusive medium nowadays. With all walks of life represented on the Web, distinct communities can be identified that form a Web of personal relationship and focus on a specific topic or theme. Different communities may include fans of a music group or supporters of a sports club, who shout out their fondness on the web; fond users of a service, say the online catalogue of a public library, who link to the library's website; or also the players of a specific online game.

Various initiatives are researching methods for automatically identifying Web communities, and their motivations are manifold (Kumar 1999). For instance, for improving search-engines and creating new services, knowledge of Web communities and their structural characteristics is of eminent importance. Where search engines used to be praised for their comprehensiveness, returning a rich set of more or less relevant Web pages to a specific query, Web users are increasingly overwhelmed by the flood of information available on the Web. The search engine Google excelled to become the number one search engine not because it returned the most results of all the search engines, but because it ranked the results according to their estimated value and thereby greatly assisted the user in finding relevant information. The technique for identifying Web communities may allow the next quantum leap in search technology. A user may then be able to better focus on a specific area of interest and prune all the other information (Flake et al. 2002). Visionary installations of this technique include advanced content filtering and the largely automatic creation of web portals, ensuring that they are always up-to-date and as complete as possible. Recommendation systems for specialised fields such as the music recommender by (Knees et al. 2005) are within reach.

Identifying Web communities enables producing opinion polls to any kind of topic at any time. The evolution of opinions can be tracked and comparisons established. One of the currently most vibrant online community spaces, the body of public weblogs, illustrates this opportunity. In the years from 1999 onwards, the public blogosphere grew at a tremendous rate. Soon a number of blog engines emerged to provide search and a variety of other services based on mining techniques. For instance Daypop[12] offers "Word Bursts" of often used words and phrases, and Technorati features ranking services services on blogs, news, books, and movies. BlogPulse[13] provides the exciting services "Trend Search" and "Conversation Tracker". The latter extracts sequences of blogs, comments and re-blogs in a single conversational trail. "Trend Search" gives the blogosphere's level of interest into a topic as it develops over time (Aschenbrenner and Miksch 2005).

By 2005 the social network of public weblogs continues its highly dynamic development. New features pop up faster than those existing can be explored, and the blogosphere is increasingly interwoven with other

[12] Daypop. www.daypop.com

[13] BlogPulse. www.blogpulse.com

emerging applications of the social software movement. More exciting applications of mining techniques can be expected from the blogosphere community, which is a great inspiration for respective mining activities in Web archives.

Mining specific communities and the evolution of the communities themselves also reflects public trends that can be used in business. Companies may choose to align their advertising campaigns according to these trends, or they may direct online advertising campaigns to specific communities. “Community pressure” may even lead to abandoning a specific product, or indicate market gaps. All in all a company’s strategic management could be informed by various conceivable analyses building on the identification of Web communities (Reid 2003).

Web communities are particularly dynamic social entities that may emerge only to quickly vanish again, or they may persist over extended periods of time. Web archives that record this offer a valuable resource for sociological research. These archives can answer questions such as the effect of a historical event on society: How many disparate communities formed around a specific event (e.g., a sports event, an election, or a political topic)? How large did they grow? and how long did they last?

Various techniques have been developed to identify Web communities. All these techniques build on the hyperlink structure of the Web, assuming that the members of the same community tend to reference more often their peers than external Web pages. In other words, identification of Web communities comes down to clustering more tightly coupled Web pages or websites together. Techniques include Bibliographic Metrics and Bipartite Cores, which build a community starting from a local set of Web pages; the HITS (hyperlink-induced topic search) Communities and the PageRank Communities algorithms, which consider all edges in the global Web graph for building a community; and the Community Algorithm, which equally works on the global Web graph as on a local sub-graph (Flake et al. 2003). The data used for community identification are basically only the links in the content of the Web pages. Specific techniques may use the text surrounding the links as a supplementary factor, but essentially no additional data than the plain Web content is used.

Experiences in community extraction show that the number of cyber-communities has consistently increased over time (Donato et al. 2004). Toyoda and Kitsuregawa analysed the evolution of Web communities in a Web archive of Japanese websites with annual crawls starting from 1999 to 2002 (Toyoda and Kitsuregawa 2003). They found that communities are relatively stable, considerably more stable than the stability of Web pages. However, although the communities as such are relatively stable, their structure changes dynamically; a majority of the identified communities is involved in merges and splits from one snapshot in the Web archive to the subsequent one. While the number of Web communities more than doubled between 1999 and 2002, the size of the individual communities was stable over that time.

These initial experiments already show the intriguing findings and the multiplicity of applications that community identification on public Web pages can produce. As the necessary data will be available in most Web archives, this branch of Web mining is a valuable addition to all Web archiving initiatives.

7.4.3 Screening Web Users

Web usage mining is increasingly employed by corporate websites and in eCommerce. It mainly builds on Web usage. On a basic level it answers questions like what parts of a website a user visited in a single session, and where she spent most time. Starting from such basic information, a number of valuable analyses can be conducted for organisational websites. Web log analysis is used for characterising the designated community of an online service (Brody and Hickman 2000), based on usage patterns the online effect of promotional activities can be gauged, or the structure of the site map can be improved. For instance, if repeatedly users view a combination of Web pages in a single session it may be convenient to provide a direct hyperlink to and from those pages. The website may even be adaptive in the sense that its Web pages adapt themselves dynamically to improve site navigation (Perkowitz and Etzioni 1997).

Usage mining can also be employed for recommender systems (Mobasher 2004) In fact, recommender system are increasingly wide spread in online shops, and many of us are likely to have encountered the kind suggestion “Users who viewed/bought this product also viewed ...”. While recommender systems mostly work with anonymous users, personalisation of services is the current buzzword that promises to revolutionise Web navigation. Thereby, user data are stored at the Web server or in cookies on the client side and reused at later sessions. It is, in fact, possible to infer various demographic data with a relatively high accuracy only based on user session data, including gender, age, income, and marital status (Murray and Durrell 2000). Such personal information may inform business policy making or may facilitate more directed advertising.

While usage mining is an increasingly important eCommerce analysis activity, there is at this time no apparent long-term value of usage data. It may be a valuable source for sociological research in the far future, but clear usage scenarios are still missing. For comparison, the longitudinal study of Web usage (Cothey 2002) took a view over a ten-month period, where longitudinal studies in other areas including Web content usually take a view over years and decades. Only (Covey 2002) takes a slightly more long-term view and also underlines the value of mining user logs over a number of different websites and organisations. Only just the fact that the value of usage data is unclear at this point of time does, however, not imply that it is worthless. This still is a young field and scholars in the future may find usage data a truly rich source.

Thus, while the long-term value of user logs may not be obvious to Web archiving initiatives tackling the broad collection of external websites today, they still bear a unique type of information that may prove indispensable in the future. Web usage data may, for example, facilitate the identification of relevant sub-trees of a website, and may thereby solve automatically what the Pandora[14] project at the National Library of Australia approaches in a manual way. Also (Chakrabarti et al. 2000) points to the value of personal references created by users intentionally or as a side product of Web browsing. Community powered Web archives such as the DACHS archive (Gross 2003) already now have the possibility for preserving this kind of data. Also community services as the one provided by Alexa[15] are in position to preserve user session data, and may inform the collection policies of Web archives or serve as a prolific data source for research.

7.4.4 Researching Networks

The previously described use cases became gradually more abstract, starting from surveying the actual content and Web usage, to analysing who visits websites. This section adds another level of abstraction and analyses the Web as an entity. The Internet is a prolific place for theoretical and mathematical research. Current research on the Internet includes self-organisation and fractal growth, graph theory, as well as game-theoretic analyses (Czumaj et al. 2002).

Probably the most exciting area in this respect currently is network theory. In fact, the Web was the trigger for a quantum leap in network theory (Bharat et al. 2003). In the late 1990s researchers found many properties of the Web to be incompatible with traditional network theory. Complex networks were formerly considered to be completely random. The link structure of the Web, however, has a considerable number of nodes, to which an

[14] Pandora project at the National Library of Australia. http://pandora.nla.gov.au/

[15] Alexa. http://www.alexa.com/

enormously high number of links point to. These nodes are called "hubs" and they reflect a property of the Web called "scale free", which basically means that already popular Web pages tend to become even more popular on the Web. Properties of the Web that allow the Web to be scale-free in the first place are its growing and dynamic nature. This was not considered by network theory until the late 1990s.

Since the discovery that the Internet defies traditional paradigms, an array of other scale-free networks has been discovered in diverse scientific areas (Albert and Barabási 2002). They include protein interaction maps in cells, social relationships, research collaborations and global trade networks. The research assumption is that the underlying mechanisms as discovered in the Web apply equally to all complex, dynamic networks, and that any findings can be transferred between these seemingly so different networks. Therefore, a number of professions including sociologists, biologists, and computer scientists follow current developments in network research with equal interest.

Currently, various initiatives are embarking on massive research efforts regarding complex networks, including the research group at the University of Notre Dame[16] and the European Commission funded project COSIN – COevolution and Self-organisation In dynamical Networks.[17] Their goal is to improve the stability, efficiency and functionality of artificial complex networks such as the Web. While dynamical networks are robust against accidental failure they are vulnerable to coordinated attacks, since the corruption of a few central nodes may disturb the whole system. Dynamical network theory may, therefore, enable new approaches to remediate hacker and virus attacks. Also, biological viruses spread in scale-free human networks, and further research on epidemic spreading may guide immunization policies in the future. A myriad of other theories and applications may emerge from dynamical networks in areas including group dynamics, linguistics, and crisis prevention in global markets. Much remains to be done until applications as mentioned earlier will materialise.

Recent research (Pennock et al. 2002) indicates that while the Web is scale-free on a high level this does not apply at a lower level; in focused communities and between peers the link distribution is less biased and not scale-free. Also the dynamic evolution of complex networks still poses many open questions.

[16] Center for Complex Network Research, University of Notre Dame. www.nd.edu/~networks

[17] COSIN – COevolution and Self-organisation In dynamical Networks. IST-2001-33555. http://www.cosin.org/

The data used for network theory research includes both Web pages as well as infrastructure data. Web archives could be the test bed for empirical research and experiments on the numerous open questions of a young and prosperous field of science.

7.4.5 Planning Network Infrastructure

As the Internet is the hinge of the modern communications infrastructure, its smooth operations are essential for business, government, and the public. Various organisations and initiatives have embarked on monitoring and analysing global network traffic, to ultimately make networks more robust and efficient. Amongst those are the Internet performance measurement and analysis (IPMA),[18] the CNRG Research Group on Internet Infrastructure,[19] the European Project SCAMPI,[20] as well as the cooperative association for internet data analysis (CAIDA).[21] Their findings and experiences focus on identifying network traffic congestion areas and ensuring that consumption is evenly spread between existing resources, as well as on projecting future network traffic development and simulating the impact of novel protocols and applications. These analyses are the basis for optimal deployment and configuration of existing network resources, and for guiding future investments in network infrastructure on both a global and a corporate level.

An example for the application of network analysis is Lumeta Corporation[22] that uses traceroute probes to explore corporate networks. Thereby, they are capable of identifying leaks or misconfigurations in the network. They can even identify and visualise organisational changes that impact on the corporate network. After mergers and acquisitions, for example, formerly separate corporate networks grow together, which can be followed in an archive containing traceroute data of the respective networks. Lumeta is a spin-off company from the Internet mapping project (Cheswick and Burch 2004) described earlier, and it continues to maintain and extend the archive of daily Internet traceroute probes that was started by the Internet Mapping Project in 1998. This database is used by various initiatives to study Internet evolution, as well as diverse other matters including graph

[18] Internet Performance Measurement and Analysis (IPMA). http://www.merit.edu/~ipma/

[19] CNRG Research Group on Internet Infrastructure. http://www.cs.cornell.edu/cnrg/

[20] SCAMPI – Scaleable Monitoring Platform for the Internet. http://www.ist-scampi.org/

[21] Cooperative Association for Internet Data Analysis (CAIDA), San Diego Supercomputing Center (SDSC). http://www.caida.org/

[22] Lumeta Corporation. http://www.lumeta.com/

theory. Indicative visualisations can be found on the archive website (Cheswick and Burch 2004), including a depiction of the changing topology of the Serbian network during the war in Yugoslavia in 1999. While it is obvious that the network took some damage during the war and some vital connections were eradicated, this time-lapse motion view shows the stability of Internet infrastructure even in face of major catastrophes.

Network analysis can this way guide policy-making and investment decisions on a corporate, national, as well as a global level. An analysis of global Internet infrastructure evolution shows that the Internet hype in the 1990s led to the installation of a massive overcapacity. Many companies failed and dropped off the Internet when the Internet bubble burst in late 2000 with continued deflation throughout 2001. The Internet economy slump can be clearly recognised in infrastructure data from that time.[23] Today the Internet infrastructure grows at a healthy rate. The rapid growth in the 1990s has stopped, however the overcapacity installed in the 1990s is still not fully used (Odlyzko 2003). It primarily grows at the periphery and thereby facilitates Internet access for a growing number of users.

Still little is known about patterns in Internet growth, but advanced research activities are working on filling the gaps. Increasingly initiatives manage to successfully combine existing approaches and knowledge Siganos et al. (2002) indicate the potential of combining various data at different levels – the network, the node, and the routing level. Also, the application of approaches from mathematical network analysis as described earlier yield a more accurate picture of infrastructure dynamics and its evolution (Vázquez et al. 2002; Siganos et al. 2002). Only a comprehensive Web archive that includes various forms of data and allows for the flexible plugging-in of novel mining techniques is able to support this kind of future research.

7.5 Conclusion

There is a myriad of conceivable Web mining applications that may be suitable for Web archives. It is the goal of this chapter to convey the breadth of the field and the opportunities at hand in order to trigger thought and pave the way for planning and design at the individual Web archiving initiative. Web archiving continues to be a highly dynamic field, and exciting new applications continue to pop up.

[23] (Broido et al. 2002), specifically Chap. 7.2.

Web archiving goes beyond technological issues and the mere collection of material from the Internet. Usage related issues need to be viewed from an organisational, functional as well as a technological perspective, and usage plays into all other functional elements of a Web archiving initiative from selection and collection, to information management and preservation. However, thorough planning and design regarding Web archive usage with a projection of possible future usage patterns by the designated community can only be done with an understanding for what is possible already now in the way of Web mining and archive usage, and an idea of the vast opportunities. Future applications need to be supported already now by an inclusive resource selection and an open system design of the Web archive.

This chapter first introduced some of the core types of data relevant for Web archiving projects. This list can no doubt be extended with more specific types and those types confined to specific initiatives for their limited availability. A full account of the possibilities and the respective requirements is impossible in the face of the diversity and the ongoing rapid development of the Internet.

Subsequently the chapter gave an impression of the possibilities in Web mining and the vastness of the field by presenting a range of research activities and Web mining projects. With some imagination it is possible to conceive of other applications future archives may provide. It has already been alluded to possible services for businesses that provide strategic data for advertisement campaign, identify emerging markets, gauge the potential for new products and thereby indicate investment opportunities. Sociology research in Web archives may identify and mitigate the digital divide with regard to geography and the varying technological opportunities and network infrastructure between regions, as well as on a societal level with mining techniques that explore the possible exclusion from the Internet of specific groups of society. As a more popular service, a Web archive may extract the "topic of the year" in the Internet and compare it with the relative coverage in traditional print media. After all, a talked-about topic in the Internet that spawns a variety of Web communities reflects the interest and opinion of the people in a more direct manner than print media.

Such opportunities trigger the vision of a future with a multiplicity of Web archive services that permeate all areas of society. Due to the still unclear requirements of such future Web mining applications, however, it is a pivotal requirement for any Web archive system to be flexible for accommodating novel mining techniques as well as new data sources. Already now Web archives should attempt to include a variety of different data as a basis for multiple, authoritative services.

As the technology environment around us changes, new forms of data will emerge that may be a valuable addition to a Web archive or that may entirely transform a Web archive. With the advent of the Semantic Web, Web archiving initiatives face novel challenges. Future technologies such as Web services and intelligent agents will be difficult to seize, even more difficult than the challenge by the deep Web as perceived today. With new approaches and tools to emerge, however, we will be able to capture aspects of the Semantic Web to the level required by the respective initiative and supported by its sphere of influence. An archive of such a semantically augmented Web sphere may offer new opportunities for Web mining and usage scenarios unheard of. An exciting future for the Internet as well as for Web archiving is wide open.

References

Aschenbrenner, A. & Miksch, S. (2005). *Blog Mining in a Corporate Environment*. Smart Agent Technologies, Research Studio. Technical Report

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics 74*

Arvidson, A., Persson, K. & Mannerheim, J. (2000). *The Kulturarw3 project - The Royal Swedish Web Archiw3e – An example of "complete" collection of web pages.* Paper presented at the 66th IFLA – International Federation of Library Associations and Institutions, Jerusalem

Barabási, A.-L. & Bonabeau, E. (2003). Scale-free networks. *Scientific American. 288*

Rekhter, Y. & Li, T. (1995). A Border Gateway Protocol 4 (BGP-4). *RFC 1771*

Bharat, K., Chang, B.-W., Henzinger, M. & Ruhl, M. (2001). *Who Links to Whom: Mining Linkage Between Web Sites*. Paper presented at the IEEE International Conference on Data Mining (ICDM'01), San Jose, California

Bhowmick, S., Keong, N. & Madria, S. (2000). *Web Schemas in WHOWEDA*. Paper presented at the ACM 3rd International Workshop on Data Warehousing and OLAP, Washington, DC

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems. 30*(1–7)

Brody, T. & Hickman, I. (2000). *Bibliometric Analysis: Mining the Social Life of an ePrint Archive*. The Open Citation Project: User studies: mining Web logs and user surveys. http://opcit.eprints.org/ijh198/

Broido, A., Nemeth, E., & Claffy, K. C., (2002). Internet Expansion, Refinement, and Churn. *European Transactions on Telecommunications 13*

Dodge, M. (2004). An atlas of cyberspace. http://www.cybergeography.com/

Chakrabarti, S., Srivastava, S., Subramanyam, M. & Tiwari, M. (2000). *Using Memex to Archive and Mine Community Web Browsing Experience*. Paper presented at the 9th International World Wide Web Conference, Amsterdam.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). *Web Mining: Information and Pattern Discovery on the World Wide Web*. Paper presented at the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Newport Beach, CA

Cothey, V. (2002). A longitudinal study of World Wide Web users' information-searching behavior. *Journal of the American Society for Information Science and Technology 53*(2). ISSN 1532-2882

Covey, D. T. (2002). Usage and Usability Assessment: Library Practices and Concerns. *CLIR Publication 105*. Digital Library Federation, Washington

Czumaj, A., Krysta, P., & Vöcking, B. (2002). Selfish Traffic Allocation for Server Farms. Paper presented at the 34th Annual ACM Symposium on Theory of Computing, Montreal, Canada

Donato, D., Laura, L., Leonardi, S., & Millozzi, S. (2004). Large scale properties of the Webgraph. *European Journal of Physics B. 38*

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). Self-Organization and Identification of Web Communities. *IEEE Computer 35*(3)

Flake, G. W., Tsioutsiouliklis, K., & Zhukov, L.(2003). *Methods for Mining Web Communities: Bibliometric, Spectral, and Flow*. In Poulovassilis, A., Levene, M. (Eds.), *Web Dynamics*. Springer, Berlin Heidelberg New York

Gross, J. (2003). *Learning by Doing: the Digital Archive for Chinese Studies (DACHS)*. Paper presented at the 3rd ECDL Workshop on Web Archives. Trondheim, Norway

Hirai, J., Raghavan, S., Garcia-Molina, H., & Paepcke, A. (2000). *Webbase: A Repository of Web Pages*. Paper presented at the 9th International World Wide Web Conference (WWW9). Amsterdam, The Netherlands. Elsevier Science

Cheswick, B. & Burch H. (2004). *Lumeta Corp.: Internet Mapping Project*. http://research.lumeta.com/ches/db/

Hallam-Baker, P. M. & Behlendorf, B. (1996). *Extended Log File Format*. W3C Working Draft, WD-logfile-960323

Inmon, W. (1992). *Building the Data Warehouse*. Wiley, New York

Kimball, R. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* Wiley, New York

Knees, P., Pampalk, E., & Widmer, G. (2005). Automatic Classification of Musical Artists based on Web-Data. *ÖGAI Journal 24*(1)

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for Emerging Cyber-Communities. *Computer Networks 31*(11)

Kushmerick, N. (2000). Wrapper Induction: Efficiency and expressiveness. *Artificial Intelligence 118*(1–2)

Leung, S., Perl, S., Stata, R., & Wiener, J. (2001). *Towards Web-scale Web Archaeology*. Research Report 174. Compaq Systems Research Center, Palo Alto, CA

Mobasher, B. (2004). Web Usage Mining and Personalization. In Singh, P.M. (Ed.), *Practical Handbook of Internet Computing*. CRC, West Palm Beach, FL, USA

Murray, D. & Durrell, K. (2000). Inferring demographic attributes of anonymous Internet users. *Lecture Notes in Artificial Intelligence 1836*. Springer, Berlin Heidelberg New York

Odlyzko, A. M. (2003). Internet traffic growth: Sources and implications. In Dingel, B., Weiershausen, W., Dutta, A. K., Sato, K.-I. (Eds.), *Optical Transmission Systems and Equipment for WDM (Wavelength-Division Multiplexing) Networking II*. SPIE (The International Society for Optical Engineering), 5247

Pennock, D., Flake, G., Lawrence, S., Glover, E., & Lee Giles, C. (2002). *Winners Don't Take All: Characterizing the Competition for Links on the Web*. Proceedings of the National Academy of Sciences *99*(8)

Perkowitz, M. & Etzioni, O. (1997). *Adaptive Sites: Automatically Learning from user Access Patterns*. Paper presented at the 6th International World Wide Web Conference, Santa Clara, CA

Phillips, M. (2003). Balanced Scorecard Initiative 49 - Collecting Australian Online Publications. Version 6. National Library of Australia

Rauber, A. & Aschenbrenner, A. (2001). Part of Our Culture is Born Digital - On Efforts to Preserve it for Future Generations. *TRANS. On-line Journal for Cultural Studies (Internet-Zeitschrift für Kulturwissenschaften) 10*. INST

Rauber, A., Aschenbrenner, A., & Witvoet, O. (2002). *Austrian On-Line Archive Processing: Analyzing Archives of the World Wide Web*. Paper presented at the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2002), Rome, Italy. Springer, Berlin Heidelberg New York

Reid, E. (2003). *Identifying a Company's Non-Customer Online Communities: a Proto-typology*. Paper presented at the IEEE Hawaiian International Conference On System Sciences (HICSS 2003), Big Island, Hawaii

Siganos, G., Faloutsos, M., & Faloutsos, C. (2002). The Evolution of the Internet: Topology and Routing. *Technical Report 65*. Carnegie Mellon University, Department of Computer Science

Toyoda, M. & Kitsuregawa, M. (2003). *Extracting Evolution of Web Communities from a Series of Web Archives*. Paper presented at the 14th ACM conference on Hypertext and Hypermedia, Nottingham, UK. ACM, New York

Vázquez, A., Pastor-Satorras, R., & Vespignani, A. (2002). Large-scale topological and dynamical properties of the Internet. *Physical Review E65(066130)*, American Physical Society

Zaiane, O. R. (1999). Resource and Knowledge Discovery from the Internet and Multimedia Repositories. PhD thesis (Simon Fraser University)