Errata

P. 3, second paragraph. The name in the first sentence should read:
"H. Xie."

P. 12, reference 10. The name should read:
"Xie, H."

P. 20, section 2.1.1.1. The last sentence should read:
"Nonetheless, due to the fact that manual pre-categorization is rather expensive, semi-supervised approaches, which diminish the need for a large number of training documents, have also been implemented (see, for example, Blum & Mitchell 1998; Liere & Tadepalli 1998; McCallum et al. 2000)."

P. 62, fifth paragraph. The name in the fourth sentence should read:
"Xie[8]."

P. 73, third paragraph. The figure at the end of the second sentence should read:
"22%."

P. 83, reference 8. The name should read:
"Hong Xie."

P. 87. Footnote should read:
"Accompanying poster text (published in conference proceedings)."

P. 93. Footnote should read:
"Accompanying poster text (published in conference proceedings)."

P. 98. Third bullet point on the page should read:
"implement a mechanism to allow identification of pages outside Renardus that users explore as a result of Renardus navigation."

P. 101, Abstract. The number in the second sentence should read:
"1003."

P. 110. Equation 3 should read:
"S = 3*STi + 2*SH + 2*SM + STe ."

P. 110, paragraph following equation 3. The second sentence should read:
"Similar co-efficients have also been derived using partial overlap (respectively): 2, 2, 2, 1 in three-digit overlap, and 2, 1, 1, 1 in two-digit overlap."

Koraljka Golub

Automated Subject Classification of Textual Web Pages, for Browsing

LUND UNIVERSITY

Department of Information Technology
Lund University, Sweden

ISBN: 91-7167-034-3
Tryckeriet i E-huset, Lund 2005

Automated Subject Classification of Textual Web Pages, for Browsing

Koraljka Golub

Automated Subject Classification of Textual Web Pages, for Browsing

Thesis for the degree of Licentiate in Philosophy,
Swedish intermediate degree between Master's and Doctoral degrees

Koraljka Golub

Department of Information Technology, Lund University
26 August 2005

Knowledge Discovery and Digital Library Research Group (KnowLib),
Digital Information Systems Group
Department of Information Technology
Lund University
http://www.it.lth.se/knowlib/

ISBN: 91-7167-034-3
ISRN: LUTEDX/TEIT-05/1031-SE

Printed in Sweden
E-huset, Lund, 2005

Abstract

With the exponential growth of the World Wide Web, automated subject classification of Web pages has become a major research issue in information and computer sciences. Organizing Web pages into a hierarchical structure for subject browsing is gaining more recognition as an important tool in information-seeking processes.

In this thesis, different automated classification approaches, focusing on organizing textual Web pages into a browsable hierarchical structure, were critically examined and compared. Three major approaches to automated subject classification have been recognized, each coming from a different research community: machine learning, information retrieval and library science. While these approaches have common research aims and a number of methods and techniques, and as such could benefit from each other, it has been shown that authors belonging to the three communities do not communicate with authors from the other two communities to a large extent. The two biggest differences between the approaches are whether they employ a vector space model (machine learning and information retrieval), and whether they make use of controlled vocabularies such as, for example, classification schemes, thesauri, or ontologies (library science).

Certain special characteristics of Web pages (e.g. metadata and structural elements such as title, headings, main text) were investigated as to how they could be best used in automated classification. The study indicated that all the structural information and metadata available in Web pages should be used in order to achieve the best automated classification results; however, the exact way of combining them proved not to be very important.

It has been claimed that well-structured, high-quality controlled vocabularies, could serve as good browsing structures. The degree and nature of subject browsing conducted by users of a large Web-based service (Renardus) was studied, using log analysis. The study showed that browsing is used to a much larger degree than searching, indicating the usefulness of browsing in such services and possibly implying the suitability of such a controlled vocabulary (Dewey Decimal Classification) for browsing.

Keywords

Subject classification, automated classification, Web page classification, text categorization, document clustering, subject browsing, structural Web-page elements, bibliographic coupling.

Contents

Preface

This thesis is a summary of papers listed below. References to the papers are made using upper-case Roman numbers associated with the papers (**I**, **II**, **III**, **IV**, **V**, **VI**).

I. Golub, K. 2005. Automated subject classification of textual Web documents. *Accepted for publication in Journal of Documentation*. Manuscript available at: http://www.it.lth.se/koraljka/Lund/publ/AC-JDoc.pdf

II. Golub, K., and Larsen, B. 2005. Different approaches to automated classification: Is there an exchange of ideas? In: *Proceedings of ISSI 2005 – the 10th International Conference of the International Society for Scientometrics and Informetrics, Stockholm, Sweden, 24-28 July 2005, Vol. 1*. P. 270-274. Also available at: http://www.it.lth.se/koraljka/Lund/publ/ISSI05.pdf

III. Koch, T., Golub, K., and Ardö, A. 2005. Users browsing behaviour in a DDC-based Web service: A log analysis. *Accepted for publication in Cataloging & Classification Quarterly*. Manuscript available at: http://www.it.lth.se/koraljka/Lund/publ/Renardus05.pdf

IV. Koch, T., Ardö, A., and Golub, K. 2004. Browsing and searching behavior in the Renardus Web service: A study based on log analysis. In: *Global reach and diverse impact: Joint Conference on Digital Libraries, Tucson, Arizona, June 7-11, 2004*. P. 378.

DOI: http://doi.acm.org/10.1145/996350.996444. Also available at: http://www.it.lth.se/koraljka/Lund/publ/JCDL04postertext.pdf

V. Koch, T., Ardö, A., and Golub, K. 2004. Log analysis of user behaviour in the Renardus Web service. In: *Human Information Behaviour & Competences for Digital Libraries: Libraries in the Digital Age, Dubrovnik and Mljet, Croatia, May 25-29, 2004*. P. 175-177. Also available at: http://www.it.lth.se/koraljka/Lund/publ/LIDA04postertext.pdf

VI. Golub, K., and Ardö, A. 2005. Importance of HTML structural elements and metadata in automated subject classification. In: *Proceedings of ECDL 2005 – the 9th European Conference on Research and Advanced Technology for Digital Libraries, Vienna, Austria, 18-23 September 2005*. P. 368-378. Also available at: http://www.it.lth.se/koraljka/Lund/publ/ECDL05.pdf

Acknowledgements

First and foremost, my deepest thanks go to my supervisors Anders Ardö and Traugott Koch. It was through their knowledge, persistence and kindness that I came to this stage of my research.

I would also like to thank Birgitta Olander for taking time out from her busy schedule to review this thesis.

Special thanks go to a number of colleagues and friends who have been giving feedback on my work: Birger Larsen, Johan Eklund, Ingo Frommholz, Repke de Vries, Jessica Lindholm, Tatjana Aparac Jelušić, Boris Badurina, Sanjica Faletar Tanacković, Martina Dragija Ivanović and Liv Fugl.

I also wish to express my gratitude to Bruce Fairfield, Goran Vukušić, Jonas Ekedahl, and Håkan Englund for proofreading my work.

Many thanks go to my colleagues from the IT Department, most of all to Suleyman Malki, for all the help and support, as well as for creating a wonderful working milieu.

Finally, I would like to express my gratitude to my family and friends, for their constant moral support and belief in me.

This research was funded by ALVIS, an EU Sixth Framework Programme, Information Society Technologies (IST-1-002068-STP) project, and The Swedish Agency for Innovation Systems (P22504-1 A).

Summary

1 Introduction

Automated subject classification has been a challenging research issue for several decades now. Major motivation has always been the high cost of manual classification. Interest has grown since the late 1990s, when using only full-text retrieval techniques in search engines became insufficient, because the number of available Web pages grew exponentially. At the same time, the library science community recognized the danger that established objectives of bibliographic systems (finding, collocating, choice, acquisition, navigation) (Svenonius 2000, p. 20-21) would get left behind, and that automated means could be a solution to preserve them (ibid., p. 30). Automated subject classification today finds its use in a wide variety of applications; apart from organizing documents into subject categories for browsing, it is used, for example, for topical harvesting, personalized routing of news articles, filtering of unwanted content for Internet browsers, and in many other tasks (see Sebastiani 2002, and Jain 1999).

In the narrower focus of this thesis is organization of textual Web pages into subject categories, for browsing. There are few research reports on automated subject classification for browsing, one of the reasons being that subject browsing has been considered under-used (Lazonder 2003, Nielsen 1997). On the other hand, it has been claimed that subject browsing is useful in a number of information-seeking situations (Koch, Day 1997; Koch, Zettergren 1999; Foskett 1996, p. 26).

Web pages have special characteristics – hyperlinks and anchors, metadata, and structural information (contained in HTML tags) – all of which could be utilized to improve automated subject classification. However, Web pages are very heterogeneous – many of them are very short, metadata provided can be inconsistent and misused, titles tend to be general ("Home page") or non-existent, structural tags are sometimes misused, etc. How to use them in automated subject classification heavily depends on characteristics of the Web-page collection at hand.

The purpose of this thesis is to carry out research in order to acquire insights into the usage of subject browsing in a large Web-based service, and to recognize, on a general level, which elements from different classification approaches, including treatment of various characteristics of Web pages, could best be utilized in automated subject classification for browsing.

The summary is structured as follows: background information on subject browsing and terminology is given in the following chapter (2 Background); problems with automated subject classification and automated subject classification approaches are discussed in Chapter 3 (Automated Classification Approaches), which is followed by an analysis of the degree to which researchers from different classification approaches exchange ideas and methods (4 Different Approaches to Automated Classification: Is There an Exchange of Ideas?); subject browsing in a large Web-based service is discussed in Chapter 5 (Subject Browsing Based on the DDC in a Large Web-Based Service); and, a study on the importance of different structural elements of Web pages in automated subject classification is given in Chapter 6 (Importance of HTML Structural Elements in Automated Subject Classification). The summary ends with final remarks and suggestions for further research (7 Concluding Remarks).

2 Background

2.1 Subject Browsing

Subject browsing in this thesis refers to seeking for information resources by examining a hierarchical tree of broader and narrower subject classes into which the resources have been classified. Web-based services offering subject browsing are many, such as those provided by quality-controlled subject gateways (e.g. Resource Discovery Network: RDN 2004; Renardus 2001), or those provided by commercial search engines (e.g. Google Directory 2005).

While it has been reported that users prefer searching to browsing (Lazonder 2003, Nielsen 1997), T. Koch and A.-S. Zettergren (1999) claim that browsing is rather useful when users are not looking for a specific information resource, when they lack experience in performing searching, and when they are not familiar with the subject and its structure and terminology. A. Foskett (1996, p. 26) says that users who are browsers, i.e. who are looking for something to catch their interest rather than answers to specific questions, form the majority of users in public libraries. He adds that it is often an item that does not fit our existing patterns of interest that proves to be the most interesting; this concept is called serendipity, "the faculty of making happy and unexpected discoveries by accident."

Subject browsing seems not to be very well supported in information services on the World Wide Web; for example, in his study on browsing strategies and implications for design of Web search engines, X. Hong (1999) reports that existing browsing features of search engines are insufficient to users. One of the possible reasons for this underdevelopment could be that people to a large extent believe that browsing is less useful. Even within the Renardus project, an initial belief about potential user requirements was that end-users preferred searching to browsing (User requirements for the broker system: Renardus Project Deliverable D1.2. 2000). After the browsing interface has been built, it was shown that browsing was much favoured (**III, IV, V**).

2.1.1 Controlled Vocabularies for Subject Browsing

Controlled vocabularies (classification schemes, thesauri, subject heading systems, ontologies) have traditionally been used in libraries, and in indexing and abstracting services, some since the 19th century. They could serve as good-quality structures for subject browsing of Web pages (esp. classification schemes), which is partly confirmed by the fact that they are already used by a number of Web-based services, especially those providing information resources for academic users (e.g. Resource Discovery Network: RDN 2004; Renardus 2001).

All these vocabularies have distinct characteristics and are consequently better suited for some applications than others. For example, subject heading systems normally do not have detailed hierarchies of terms (exception: Medical Subject Headings), while classification schemes consist of hierarchically structured groups of classes. Since in classification schemes similar documents are grouped together into classes and relationships between the classes are established, they are better suited for subject

browsing than other controlled vocabularies (Vizine-Goetz 1996; Koch, Zettergren 1999; see also Soergel 2004). Different classification schemes have different characteristics; for subject browsing the following are important: the bigger the collection, the more depth should the hierarchy contain; hierarchically flat schemes are not effective for browsing; classes should contain more than just one or two documents (Schwartz 2001, p. 48). Search-engine directories and other homegrown schemes on the Web, "…even those with well-developed terminological policies such as Yahoo… suffer from a lack of understanding of principles of classification design and development. The larger the collection grows, the more confusing and overwhelming a poorly designed hierarchy becomes…" (ibid., p. 76). For these reasons it was decided to study user behaviour in a Web-based service that is based on a classification scheme, which has been used and maintained in libraries for more than a century now – Dewey Decimal Classification (DDC) (Dewey services 2005) (**III, IV, V**).

2.2 Classification and Automated Classification: Terminology

The term "classification" is in this thesis used as defined by Chan (1994, p. 259): it involves grouping documents that have a property in common, further sub-grouping of documents based on finer properties, and establishing relationships between them.

The term "automated subject classification" (in further text: automated classification) is in this thesis used to denote machine-based organization of subject-related information objects. Certain human intellectual processes are replaced by, for example, statistical and computational linguistics techniques.

There are three major approaches to automated classification (**I**), the most frequently used one being text categorization (coming from machine-learning community), followed by document clustering (information retrieval), and document classification (library science). The terms "text categorization" and "document clustering" are chosen because they tend to be the most frequently used terms in the literature of the corresponding communities; "document classification" was chosen for the thesis, in order to consistently distinguish between the three approaches.

3 Automated Classification Approaches

In **I**, different approaches to automated classification have been reviewed. There are three major approaches:

1) Text categorization, which is a machine-learning approach, also applying information retrieval methods. It consists of three main parts. The first part involves manual categorization of a number of documents (called training documents) to pre-defined categories. By learning the characteristics of those documents (second part), the automated categorization of new documents takes place (third part). In the machine-learning terminology, text categorization is known as supervised learning, since the process is "supervised" by learning categories' characteristics from manually categorized documents.
2) Document clustering, which is an information-retrieval approach. Unlike text categorization, it does not involve pre-defined categories or training documents and is thus called unsupervised. In this approach clusters and, to a limited degree, relationships between the clusters, are derived automatically from the documents to be clustered, and the documents are subsequently assigned to those clusters.
3) Document classification, which is a library science approach. It involves an intelectually created controlled vocabulary (most often a classification scheme). Documents are classified into classes of the classification scheme, using simple algorithms.

Major similarities between the three main approaches include document pre-processing, and utilization of Web-specific document characteristics. Major differences are in algorithms applied, employment (or not) of the vector space model and of controlled vocabularies. In the context of subject browsing, the most important difference between the three main approaches to automated classification is whether or not they use a controlled vocabulary, and if so, how suitable that vocabulary is for subject browsing. In document classification, classification schemes are usually well structured for browsing, and names used for classes have been carefully chosen. In text categorization, categories are manually constructed, but often only few categories with one or two hierarchical levels are used in experiments, each consequently containing an "unbrowsable" number of documents. In document clustering, clusters, relationships between them, and their names are automatically produced. Labelling of the clusters is a major problem of the approach, and relationships between the clusters, such as those of equivalence, associative, and hierarchical relationships, are even more difficult

to automatically derive (Svenonius 2002, p.168); as put by H. Chen and S. Dumais (2000, p.2), "[a]utomatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand". Another problem with document clustering is that the structure of clusters and their names change as new documents are added to the collection; such instability in Web-based services and digital libraries is not user-friendly. Thus, as T. Koch and A. Zettergren (1999) suggested, document clustering is better suited for organizing Web search engine results.

An emerging approach, in this thesis referred to as mixed approach, is one in which controlled vocabularies are employed in combination with text categorization and document clustering algorithms. The emergence of the mixed approach demonstrates the potentials for utilizing ideas and methods from another community's approach.

Several problems with automated classification in general have been identified (**I**). As E. Svenonius (2000, p.46-49) claims, automating subject determination belongs to logical positivism: a subject is considered to be a string occurring above a certain frequency, which is not a stop word, and/or is found in a given location (e.g. title), or, in clustering algorithms, inferences are made such as "if document A is on subject X, then if document B is sufficiently similar to document A (above a certain threshold), then document B is on that subject." Evaluation is a major issue in automated classification. The problem of deriving the correct interpretation of a document's subject matter has been much discussed among library scientists, while much less so in machine learning and information retrieval communities. It has been reported that different people, whether users or subject indexers, would assign different subject terms or classes to the same document. Studies on inter-indexer and intra-indexer consistency also report generally low indexer consistency (Olson, Boll 2001, p. 99-101), which has to do, among other things, with the policy of indexing (e.g. higher specificity and exhaustivity lead to lower consistency), and the size of the controlled vocabulary used (the bigger the vocabulary, the more choices there are to choose from).

4 Different Approaches to Automated Classification: Is There an Exchange of Ideas?

As discussed in the previous chapter and in paper **I**, the three major approaches share a number of common ideas and methods, and the fact that they could benefit from each other is reflected in the emergence of the mixed approach. In order to determine to what a degree the three communities

explore other communities' ideas, methods, or findings, direct links (do authors from one community cite authors from another) and indirect links (using bibliographic coupling of papers) were studied (**II**).

The sample consisted of 148 papers on automated classification of textual Web pages; 63 papers were from the information retrieval (IR) community, 52 from machine learning (ML) and 33 from library science (LS). The size of the sample is limited by the small number of LS papers on that topic (while there are many ML papers, for example). Also, the library science set of papers includes two subgroups, one "pure" library science subgroup, and the other with papers using either IR or ML approach, but also applying intellectually created vocabularies. Not having any formal criteria, e.g. distinct channels of publication for each community, every paper had to be at least partially read in order to be assigned to the corresponding community.

The hypothesis was that the three different communities do not communicate with each other to a large extent. It was found that absence of communication was especially the case for the LS community, whereas the ML and IR community exchange ideas and methods to a certain degree, although they also form distinct groupings. Those papers from the subgroup of LS coming from ML or IR but using an intellectually created vocabulary coupled with ML and/or IR, and not with other, "pure" LS papers.

Further research, based on a bigger sample, is needed to determine why direct and indirect links are lacking between LS and the other two communities, in spite of appearance of ML and IR papers that employ controlled vocabularies.

5 Subject Browsing Based on DDC in a Large Web-Based Service

In order to investigate subject browsing behaviour, a study of a large Web-based service was conducted (**III, IV, V**), where browsing is provided using a well-established classification scheme. The service chosen was Renardus (Renardus 2001), which provides integrated searching and browsing access to quality-controlled Web resources from major European subject gateways.

The main navigation feature in Renardus is subject browsing through the DDC, with several browsing-support features such as the graphical fish-eye display, search entry into the browsing pages, and merging the resource descriptions from all related collections. Searching options are also well supported, allowing combinations of terms and search fields and providing options to limit searches in a number of different ways.

A frequently applied methodology for studying user information behaviour on the World Wide Web is log analysis. The authors (**III, IV, V**) have also chosen this approach, as it has several advantages: users do not need to be directly involved in the study, a picture of user behaviour is captured in non-invasive conditions, and every activity inside the system can be tracked. Our own software for log analysis has been developed, since existing software packages did not support all the needed tasks. All entries in the log file were grouped into user sessions. A user session was heuristically defined as containing all entries coming from the same IP address and a time gap of less than one hour to the prior entry from the same IP-number.

The research aimed at studying: the unsupervised usage behaviour of all Renardus users, complementing the initial Renardus user enquiry; detailed usage patterns (quantitative/qualitative, paths through the system); the balance between browsing, searching and mixed activities; typical sequences of user activities and transition probabilities in a session, especially in traversing the hierarchical DDC browsing structure; the degree of usage of the browsing support features; and typical entry points, referring sites, points of failure and exit points.

In contrast to common belief (Lazonder 2003, Nielsen 1997), our study clearly indicates that browsing as an information-seeking activity is highly used, given proper conditions. About 80% of all activities in Renardus are browsing activities. A contributing reason to that dominance is the fact that a very high percentage (71%) of the users are referred from search engines directly to a browsing page in Renardus. The layout of the home page "invites" browsing, which contributes to the fact that also users starting at the home page (22%) predominantly use the browsing part of the service. The browsing support features are also heavily used, most of all graphical overview and search entry to browsing pages.

People starting at the homepage show almost twice as many activities per session, and use the non-browsing features three to five times as often. Their share of the browsing activities is smaller, but they primarily engage in the long sequences of browsing activities (8 and longer) and employ more different types of browsing and more different types of other activities in a session. The home page starters are a minority but use the service elaborately, in a way the system designers have imagined and intended.

The DDC directory browsing is the single clearly dominating activity in Renardus (60%). Two-thirds of it is done in unbroken directory browsing sequences. There is a surprising average and total length of such browsing sequences – while the majority limit themselves to about 10 such steps, long

unbroken sequences of up to 86 steps in the DDC directory trees were found.

The study also indicates that a thorough log analysis can indeed provide a deeper understanding of user behaviour and service performance. Being an unobtrusive means of capturing unsupervised usage and offering a complete and detailed picture of user activities, it can reveal quantitatively comprehensive results.

6 Importance of HTML Structural Elements in Automated Subject Classification

This study (**VI**) is based on an automated classification approach (Ardö, Koch 1999) that has been developed within the DESIRE project (DESIRE 2000) as part of the subject gateway Engineering Electronic Library (Engineering Electronic Library 2003). With the overall purpose of improving the classification algorithm, the aim was to determine the importance of distinguishing between different parts of a Web page. Significance of four elements was studied: title, headings, metadata, and main text. The hypothesis was that best automated classification results are achieved when appropriate significance indicators are assigned to the structural elements and metadata. A data collection that was used consisted of some 1000 Web pages in engineering, to which Ei classes (Ei thesaurus 1995) have been manually assigned.

The significance indicators were derived using several different methods: precision and recall, partial precision and recall, semantic distance and multiple regression. Precision is in the context of automated classification defined as the share of correctly assigned classes in all automatically assigned classes. Recall is defined as the share of correctly assigned classes in all manually assigned ones. The Ei classification scheme has a solid hierarchical structure, thus allowing for a rather credible test on partial overlap. Three different levels of overlap were tested:

1. total overlap, e.g. if the class "932.2.1." is the correct one, than the one automatically assigned needs to look exactly the same;
2. partial overlap when the first three digits are identical, e.g. "932.5" and "932.2."; and,
3. partial overlap when the first two digits are identical, e.g. "932" and "933".

Semantic distance is a numerical representation of the difference in meaning between two classes. Since Ei classification scheme embodies a well-

developed network of hierarchical relationships between classes, it is rather straightforward to use the semantic distance measure. The following measures were used: 4, when the classes differ already in the first digit (e.g. 6 vs. 901); 2, when the classes differ already in the second digit (e.g. 932.3 vs. 901.3); 1, when the classes differ in the third digit (e.g. 674.1 vs. 672); and 0.5, when the classes differ in the fourth digit (e.g. 674.1 vs. 674).

Multiple regression served as another method against which results could be compared. It was used in a rather simplified way: scores assigned by each of the algorithms were taken as independent variables, while the final score represented the dependent variable. It was set to either 1000 or 0, denoting correct and incorrect classes respectively. Derived regression coefficients in our case represented the significance indicators.

Different significance indicators, derived using those four methods, were tested against the baseline algorithm (where all the indicators are considered equal). The results have shown that using all structural elements and metadata is necessary since not all of them occur on every page. However, the exact way of combining the significance indicators turned out not to be highly important: the best combination of significance indicators (gives big significance to classes that were assigned based on the title) is 3% better than the baseline.

Reasons for such results need to be further investigated. One could guess that this is due to the fact that the Web pages in our data collection were rather heterogeneous; on the other hand, they were selected by librarians for end users of an operational service, and as such they might indicate what such Web-page collections are like. Apart from heterogeneity, the problem could be that metadata were abused, and that certain tags were misused (e.g. instead of using appropriate tags for making text bold, one used a headings tag, which has the same effect on the screen).

7 Concluding Remarks

While communities with different automated classification approaches do not communicate with each other to a large extent, a mixed approach has been recognized to be emerging. In order to provide good browsing structures as results of their complex and much-researched automated classification algorithms, text categorization and document clustering communities would need to employ suitable controlled vocabularies.

All the communites have made some effort into exploiting different characteristics of Web pages as to how they could be utilized in automated

classification. The study has shown that text coming from all the four parts of a Web page (title, headings, main text, metadata) should be included in the process, while the difference as to how the four elements are combined (i.e., they all seem to have about the same significance about the content) does not make a very big difference.

While subject browsing was shown to be useful in a large Web-based service, it needs to be further investigated to determine to what degree it is suitable for different users' tasks. More research needs to be done on controlled vocabularies for browsing in the electronic environment, as well as their suitability for automated classification.

References

1. Ardö, A., and Koch, T. 1999. Automatic classification applied to the full-text Internet documents in a robot-generated subject index. In: Online Information 99, Proceedings of the 23rd International Online Information Meeting, London. P. 239-246.

2. Chan, L. M. 1994. Cataloging and classification: An introduction. New York: McGraw-Hill.

3. Chen, H., and Dumais, S. T. 2000. Bringing order to the web: Automatically categorizing search results. In: Proceedings of CHI-00, ACM International Conference on Human Factors in Computing Systems, Den Haag, NL. New York: ACM Press. P. 145-152.

4. DESIRE : Development of a European Service for Information on Research and Education. 2000. http://www.desire.org/

5. Dewey services. 2005. http://www.oclc.org/dewey/

6. Ei thesaurus. 1995. 2nd ed. Hoboken, NJ: Engineering Information.

7. Engineering Electronic Library. 2003. http://eels.lub.lu.se/

8. Foskett, A. C. 1996. The subject approach to information. London: Library Assocation Publishing.

9. Google Directory. 2005. http://www.google.com/dirhp

10. Hong, X. 1999. Web browsing: Current and desired capabilities. In: 20th Annual National Online Meeting, 18-20 May 1999, New York, NY, US. P. 523-37.

11. Jain, A. K., Murty, M. N., and Flynn, P. J. 1999. Data clustering: A review. ACM Computing Surveys 31(3): 264-323.

12. Koch, T., and Day, M. 1997. The role of classification schemes in Internet resource description and discovery: EU Project DESIRE, Deliverable D3.2.3. http://www.lub.lu.se/desire/radar/reports/D3.2.3/

13. Koch, T., and Zettergren, A.-S. 1999. Provide browsing in subject gateways using classification schemes. http://www.lub.lu.se/desire/handbook/class.html

14. Lazonder, A. W. 2003. Principles for designing web searching instruction. Education and Information Technologies 8(June): 179–193.

15. Nielsen, J. 1997. Search and you may find. Jakob Nielsen's Alertbox for July 15, 1997. http://www.useit.com/alertbox/9707b.html

16. Olson, H. A., and Boll, J. J. 2001. Subject analysis in online catalogs. 2nd ed. Englewood, CO: Libraries Unlimited.

17. Renardus. 2001. http://www.renardus.org

18. Resource Discovery Network: RDN. 2004. http://www.rdn.ac.uk

19. Schwartz, C. 2001. Sorting out the Web: Approaches to subject access. Westport, CT: Ablex.

20. Sebastiani, F. 2002. Machine learning in automated text categorization. ACM Computing Surveys 34 (1): 1–47.

21. Soergel, D. et al. 2004: Reengineering thesauri for new applications : The AGROVOC example. Journal of Digital Information 4 (4). http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/

22. Svenonius, E. 2000. The intellectual foundations of information organization. Cambridge, MA: MIT Press.

23. User requirements for the broker system. 2000. Renardus Project Deliverable D1.2. 2000. http://www.renardus.org/about_us/deliverables/d1_2/D1_2_final.pdf

24. Vizine-Goetz, D. 1996. Using library classification schemes for Internet resources. OCLC Internet Cataloging Project Colloquium. http://staff.oclc.org/~vizine/Intercat/vizine-goetz.htm

All electronic resources have been accessed 20 June 2005.

Papers

I

Automated Subject Classification of Textual Web Documents

1 Introduction

Classification is, to the purpose of this paper, defined as "…the multistage process of deciding on a property or characteristic of interest, distinguishing things or objects that possess that property from those which lack it, and grouping things or objects that have the property or characteristic in common into a class. Other essential aspects of classification are establishing relationships among classes and making distinctions within classes to arrive at subclasses and finer divisions" (Chan 1994, p.259). *Automated subject classification* (in further text: automated classification) denotes machine-based organization of related information objects into topically related groups. In this process human intellectual processes are replaced by, for example, statistical and computational linguistics techniques. In the literature on automated classification, the terms *automatic* and *automated* are both used. Here the term *automated* is chosen because it more directly implies that the process is machine-based.

Automated classification has been a challenging research issue for several decades now. Major motivation has been the high cost of manual classification. Interest has grown rapidly since 1997, when search engines couldn't do with just text retrieval techniques, because the number of available documents grew exponentially. Due to the ever-increasing number

of documents, there is a danger that recognized objectives of bibliographic systems (Svenonius 2000, p.20-21) would get left behind; automated means could be a solution to preserve them (ibid., p.30). Automated classification of text finds its use in a wide variety of applications, such as: organizing documents into subject categories for topical browsing, including grouping search results by subject; topical harvesting; personalized routing of news articles; filtering of unwanted content for Internet browsers; and many others (see Sebastiani 2002, and Jain, Murty & Flynn 1999).

In the narrower focus of this paper is automated classification of textual Web documents into subject categories for browsing. Web documents have specific characteristics such as hyperlinks and anchors, metadata, and structural information, all of which could serve as complementary features to improve automated classification. On the other hand, they are rather heterogeneous; many of them contain little text, metadata provided are sparse and can be misused, structural tags can also be misused, and titles can be general ("Home page", "Untitled Document"). *Browsing* in this paper refers to seeking for documents via a hierarchical structure of subject classes into which documents had been classified. Research has shown that people find browsing useful in a number of information-seeking situations, such as: when not looking for a specific item (Koch & Zettergren 1999), when one is inexperienced in searching (ibid.), or unfamiliar with the subject in question and its terminology or structure (Schwartz 2001, p.76).

In the literature, terms such as classification, categorization and clustering are used to represent different approaches. In their broadest sense these terms could be considered synonymous, which is probably one of the reasons why they are interchangeably used in the literature, even within the same research communities. For example, A. Hartigan (1996, p.2) says: "The term cluster analysis is used most commonly to describe the work in this book, but I much prefer the term classification..." Or: "…classification or categorization is the task of assigning objects from a universe to two or more classes or categories" (Manning & Schütze 1999, p.575).

In this paper terms *text categorization* and *document clustering* are chosen because they tend to be the prevalent terms in the literature of the corresponding communities. *Document classification* and *mixed approach* are used in order to consistently distinguish between the four approaches. Descriptions of the approaches are given below:

1. *Text categorization* is a machine-learning approach, in which also information retrieval methods are applied. It consists of three main parts: categorizing a number of documents to pre-defined categories, learning the characteristics of those documents, and categorizing new documents.

In the machine-learning terminology, text categorization is known as supervised learning, since the process is "supervised" by learning categories' characteristics from manually categorized documents.

2. *Document clustering* is an information-retrieval approach. Unlike text categorization, it does not involve pre-defined categories or training documents and is thus called unsupervised. In this approach the clusters and, to a limited degree, relationships between clusters are derived automatically from the documents to be clustered, and the documents are subsequently assigned to those clusters.
3. *Document classification* in this paper stands for a library science approach. It involves an intellectually created controlled vocabulary (such as classification schemes), into classes of which documents are classified. Controlled vocabularies have been developed and used in libraries and in indexing and abstracting services, some since the end of the 19th century.
4. Sometimes methods from text categorization or document clustering are used together with controlled vocabularies. In the paper such as approach is referred to as a *mixed approach*.

To the author's knowledge no review paper on automated text classification attempted to discuss more than one community's approach. Individual approaches of text categorization, (document) clustering and document classification have been analysed by F. Sebastiani (2002), A. Jain, M. Murty & P. Flynn (1999), and E. Toth (2002), respectively.

This paper deals with all the approaches, from an integrated perspective. It is not aimed at detailed descriptions of approaches, since they are given in the above-mentioned reviews. Nor does it attempt to be comprehensive and all-inclusive. It aims to point to similarities or differences as well as problems with the existing approaches. In what aspects and to what degree are today's approaches to automated classification comparable? To what degree can the process of subject classification really be automated, with the tools available today? What are the remaining challenges? These are the questions touched upon in the paper.

The paper is laid out as follows: explorations of individual approaches as to their special features (description, differences, evaluation), application and employment of characteristics of Web pages are given in the second section (2 Approaches to automated classification), followed by a discussion (3 Discussion).

2 Approaches to automated classification

2.1 Text categorization

2.1.1 Special features

2.1.1.1 Description of features

Text categorization is a machine-learning approach, which has also adopted some features from information retrieval (see below).

The process of text categorization consists of three main parts:

1. The first part involves manual categorization of a number of documents to pre-defined categories. Each document is represented by a vector of terms. (The vector space model comes from information retrieval). These documents are called training documents because, based on those documents, characteristics of categories they belong to are learnt.
2. By learning the characteristics of training documents, for each category a program called classifier is constructed. After the classifiers have been created, and before automated categorization of new documents takes place, classifiers are tested with a set of so-called test documents, which were not used in the first step.
3. The third part consists of applying the classifier to new documents.

In the literature, text categorization is known as *supervised* learning, since the process is "supervised" by learning from manually pre-categorized documents. As opposed to text categorization, clustering is known as an unsupervised approach, because it does not involve manually pre-clustered documents to learn from. Nonetheless, due to the fact that manual pre-categorization is rather expensive, semi-supervised approaches, which diminish the need for a large number of training documents, referred to as semi-supervised ones, have also been implemented (see, for example, Blum & Mitchell 1998; Liere & Tadepalli 1998; McCallum et al. 2000).

2.1.1.2 Differences within the approach

A major difference among text categorization approaches is in how classifiers are built. They can be based on Bayesian probabilistic learning, decision tree learning, artificial neural networks, genetic algorithms or instance-based learning – for explanation of those, see, for example, T. Mitchell 1997. There have also been attempts of classifier committees (or metaclassifiers), in which

results of a number of different classifiers are combined to decide on a category (e.g. Liere & Tadepalli 1998). One also needs to mention that not all algorithms used in text categorization are based on machine learning. For example, Rocchio is actually an information retrieval classifier (Rocchio 1971), and WORD (Yang 1999) is a non-learning algorithm, invented to enable comparison of learning classifiers' categorization accuracy. Comparisons of learning algorithms can be found in H. Schütze, D. Hull & J. Pedersen (1995), Y. Li & A. Jain (1998), Y. Yang (1999), or F. Sebastiani (2002).

Another difference within the text categorization approach is in the document pre-processing and indexing part, where documents are represented as vectors of term weights. Computing the term weights can be based on a variety of heuristic principles. Different terms can be extracted for vector representation (single words, phrases, stemmed words etc.), also based on different principles; characteristics of Web documents, such as mark-up for emphasized terms and links to other documents, are often experimented with (see, for example, Gövert, Lalmas & Fuhr 1999). The number of terms per document needs to be reduced not only for indexing the document with most representative terms, but also for computing reasons. This is called dimensionality reduction of the term space. Dimensionality reduction methods could include removal of non-informative terms (not only stop words); also, taking only parts of the Web document, its snippet or summary (Mladenic & Grobelnik 2003), has been explored. For an example of a complex document representation approach, a word clustering one, see R. Bekkerman et al. (2003); for another example, based on latent semantic analysis, see L. Cai & T. Hofmann (2003).

Several researches have explored how hierarchical structure of categories into which documents are to be categorized could influence the categorization performance. D. Koller & M. Sahami (1997) used a Bayesian classifier at each node of the classification hierarchy and employed a feature selection method to find a set of discriminating features (i.e., words) for each node. They showed that, in comparison to a flat approach, using hierarchical structure could improve classification performance. Similar improvements were reported by A. McCallum et al. (1998), S. Dumais & H. Chen (2000), and M. Ruiz & P. Srinivasan (1999).

2.1.1.3 Evaluation methods

Various measures are used to evaluate different aspects of text categorization performance (Yang 1999). Effectiveness, the degree to which correct categorization decisions have been made, is often evaluated using

performance measures from information retrieval, such as precision (correct positives/predicted positives) and recall (correct positives/actual positives). Efficiency can also be evaluated, in terms of computing time spent on different parts of the process. There are other evaluation measures, and new are being developed such as those that take into account degrees to which a document was wrongly categorized (Dumais, Lewis & Sebastiani 2001; Sun, Lim & Ng 2001). For more on evaluation measures in text categorization, see F. Sebastiani (2002, p.32-39). Evaluation in text categorization normally does not involve subject experts or users.

Y. Yang (1999) claims that the most serious problem in text categorization evaluations is the lack of standard data collections and shows how different versions of the same collection have a strong impact on the performance, and other versions do not. Some of the data collections used by the text categorization community are: Reuters (Reuters-21578 2004), which contains newswire stories classified under categories related to economics; OHSUMED (Hersh et al. 1994), which contains abstracts from medical journals categorized under Medical Subject Headings (MeSH); the U.S. Patent database in which patents are categorized into the U.S. Patent Classification System; 20 Newsgroups DataSet (20 Newsgroups DataSet 1998), which contains about 20000 postings to Usenet newsgroups belonging to 20 different news groups. For Web documents there is WebKB (WebKB 2001), Cora (McCallum et al. 1999), and samples from directories of Web documents such as Yahoo! (Yahoo! Directory 2005). All these collections have a different number of categories and hierarchical levels. There seems to be a tendency to conduct experiments on a relatively small number of categories with few hierarchical levels, which is usually not suitable for subject browsing tasks.

2.1.2 Characteristics of Web pages

A number of issues related to categorization of textual Web documents have been dealt with in the literature. Hypertext-specific characteristics such as hyperlinks, HTML tags and metadata have all been explored.

Y. Yang, S. Slattery, R. Ghani (2002) have defined five hypertext regularities of Web document collections, which need to be recognized in order to chose an appropriate text categorization approach: (1) no hypertext regularity; in which case standard classifiers for text are used; (2) encyclopaedia regularity, when documents with a certain category label only link to documents with the same category label, in which case the text of each document could be augmented with the text of its neighbours; (3) co-

referencing regularity, when neighbouring documents have a common topic; in which case the text of each document can be augmented with the text of its neighbours, but text from the neighbours should be marked (e.g. prefixed with a tag); (4) preclassified regularity, when a single document contains hyperlinks to documents with the same topic, in which case it is sufficient to represent each page with names of the pages it links with; and, (5) metadata regularity, when there are either external sources of metadata for the documents on the Web, in which case we extract the metadata and look for features that relate documents being categorized, or metadata are contained within the META, ALT and TITLE tags. Several other papers discuss characteristics of document collections to be categorized. S. Chakrabarti, B. Dom & P. Indyk (1998) showed that including documents that cite, or are cited by the document being categorized, as if they were local terms, performed worse than when those documents were not considered. They achieved improved results applying a more complex approach with refining the class distribution of the document being classified, in which both the local text of a document and the distribution of the estimated classes of other documents in its neighbourhood, were used. S. Slattery and M. Craven (2000) showed how discovering regularities, such as words occurring on target pages and on other pages related by hyperlinks, in both training and test document sets could improve categorization accuracy. M. Fisher & R. Everson (2003) found out that link information could be useful if the document collection had a sufficiently high link density and links were of sufficiently high quality. They introduced a frequency-based method for selecting the most useful citations from a document collection.

A. Blum & T. Mitchell (1998) compared two approaches, one based on full-text, and one based on anchor words, and found out that anchor words alone were slightly less powerful than the full-text alone, and that the combination of the two was best. E. Glover et al. (2002) reported that the text in citing documents close to the citation often has greater discriminative and descriptive power than the text in the target document. Similarly, A. Attardi, A. Gulli & F. Sebastiani (1999) used information from the context where a URL that refers to that document appears and got encouraging results. J. Fürnkranz (1999) included words that occurred in nearby headings and in the same paragraph as anchor-text, which yielded better results than using the full-text alone. In his later study (2002) he used portions of texts from all pages that point to the target page: the anchor text, the headings that structurally precede it, the text of the paragraph in which it occurs, and a set of linguistic phrases that capture syntactic role of the anchor text in this paragraph. Headings and anchor text seemed to be most useful.

In regards to metadata, R. Ghani, S. Slattery & Y. Yang (2001) reported that metadata could be very useful for improving classification accuracy.

2.1.3 Application

Text categorization is the most frequently used approach to automated classification. While a large portion of research is aimed at improving algorithm performance, it has been applied in operative information systems, such as Cora (McCallum et al. 2000), NorthernLight (Dumais, Lewis & Sebastiani 2002, p.69-70), and the Thunderstone Web Site Catalog (Thunderstone Web Site Catalog 2005). However, detailed information about approaches used in commercial directories is mostly not available, due to their proprietary nature (Pierre 2001, p.9). There are other examples of applying machine-learning techniques to Web pages and categorizing them into browsable structures. D. Mladenic (1998) and Y. Labrou & T. Finin (1999) used the Yahoo! directory (Yahoo! Directory 2005). J. Pierre (2001) categorized Web pages into industry categories, although he used only top-level categories of North American Industrial Classification System.

Apart from organizing Web pages into categories, text categorization has been applied for categorizing Web search engine results (see, for example, Chen & Dumais 2000; Sahami, Yusufali & Baldonado 1998). It also finds its application in document filtering, word sense disambiguation, speech categorization, multimedia document categorization, language identification, text genre identification, and automated essay grading (Sebastiani 2002, p.5).

2.1.4 Summary

Text categorization is a machine-learning approach, with the vector-space model and evaluation measures borrowed from information retrieval. Characteristics of pre-defined categories are learnt from manually categorized documents.

Within text categorization, approaches can differ in several aspects. Classifiers can be based on different machine-learning algorithms. Different methods are applied to represent documents as vectors of term weights. Different evaluation measures and data collections are used. The potential added value of Web document characteristics, which have been compared and experimented with, are, for example, anchor words, headings words, text near the URL for the target document, inclusion of linked document's text as being local. When deciding which methods to use, one

needs to determine which characteristics are common to the documents to be categorized; for example, augmenting the document to be classified with the text of its neighbours will yield good results only if the source and the neighbours are related enough.

Text categorization is the most widespread approach to automated classification, with a lot of experiments being conducted under controlled conditions. There seems to be a tendency to use a small number of categories with few hierarchical levels, which is usually not suitable for subject browsing tasks. Several examples of its application in operative information systems exist.

2.2 Document clustering

2.2.1 Special features

2.2.1.1 Description of features

Document clustering is an information retrieval approach. As opposed to text categorization, it does not involve manually pre-categorized documents to learn from, and is thus known as an unsupervised approach.

The process of document clustering involves two main steps:

1. Documents to be clustered are represented by vectors, which are then compared to each other using similarity measures. Like in text categorization, different principles can be applied at this stage to derive vectors (which words or terms to use, how to extract them, which weights to assign based on what etc.). Also, different similarity measures can be used, the most frequent one probably being the cosine measure.
2. In the following step, documents are grouped into clusters using clustering algorithms. Two different types of clusters can be constructed: partitional (or flat), and hierarchical.

Partitional algorithms determine all clusters at once. A usual example is K-means, in which first a k number of clusters are randomly generated; when new documents are assigned to the nearest centroid (centre of a cluster), centroids for clusters need to be re-computed.

In hierarchical clustering, a hierarchy of clusters is built. Often agglomerative algorithms are used: first, each document is viewed as an individual cluster; then, the algorithm finds the most similar pair of clusters and merges them. Similarity between documents can be calculated in a number of ways. For example, it can be defined as the maximum similarity between any two individuals, one from each of the two groups (single-

linkage), as the minimum similarity (complete-linkage), or as the average similarity (group-average linkage).

For a review of different clustering algorithms, see A. Jain, Murty & Flynn (1999), E. Rasmussen (1992), and D. Fasulo (1999).

Another approach to document clustering is self-organizing maps (SOMs). SOMs are a data visualisation technique, based on unsupervised artificial neural networks, that transform high-dimensional data into (usually) two-dimensional representation of clusters. For a detailed overview of SOMs, see T. Kohonen (2001). There are many research examples of visualization for browsing using SOMs (see, for example, Heuser, Babanine & Rosenstiel 1998; Poincot, Lesteven & Murtagh 1998; Rauber & Merkl 1999; Goren-Bar et al. 2000; Schweighofer, Rauber & Dittenbach 2001; Yang, Chen & Hong 2003; Dittenbach, Berger & Merkl 2004).

2.2.1.2 Differences within the approach

A major difference within the document clustering community is in algorithms (cf. the above section). While previous research showed that agglomerative algorithms performed better than partitional ones, some studies indicate the opposite. M. Steinbach, G. Karypis & V. Kumar (2000) compared agglomerative hierarchical clustering and K-means clustering and showed that K-means is at least as good as agglomerative hierarchical clustering. Y. Zhao & G. Karyps (2002) evaluated different partitional and agglomerative approaches and showed that partitional algorithms always lead to better clustering solutions than agglomerative algorithms. In addition, they presented a new type of clustering algorithms called constrained agglomerative algorithms that combined the features of both partitional and agglomerative algorithms. This solution gave better results than agglomerative or partitional algorithms alone. For a comparison of hierarchical clustering algorithms, and added value of some linguistics features, see V. Hatzivassiloglou, L. Gravano & A. Maganti (2000). Different enhancements to algorithms have been proposed (see, for example, Liu et al. 2002, Mandhani, Joshi & Kummamuru 2003; Slonim, Friedman & Tishby 2003).

Since in document clustering (including SOMs) clusters and their labels are produced automatically, deriving the labels is a major research challenge. In an early example of automatically derived clusters (Garfield, Malin & Small 1975), which were based on citation patterns, labels were assigned manually. Today a common heuristics is to extract between five and ten of the most frequent terms in the centroid vector, then to drop stop-words and perform stemming, and choose the term which is most frequent in all documents of the cluster. A more complex approach to labelling is given

by E. Glover et al. (2003). They used an algorithm to predict "parent, self, and child terms"; self terms were assigned as clusters' labels, while parent and children terms were used to correctly position clusters in the cluster collection.

Another problem in document clustering is how to deal with large document collections. According to Jain, Murty & Flynn (1999, p.316), only the K-means algorithm and SOMs, have been tested on large data sets. An example of an approach dealing with large data sets and high dimensional spaces was presented by T. Haveliwala, A. Gionis & P. Indyk (2000), who developed a technique they managed to apply to 20 million URLs.

2.2.1.3 Evaluation methods

Similarly to text categorization, there are many evaluation measures (e.g. precision and recall), and evaluation normally does not include subject experts or users.

Data collections often used are fetched from TREC (TREC : Text REtrieval Conference 2004). In the development stage is the INEX initiative project (INitiative for the Evaluation of XML Retrieval 2004), within which a large data collection of XML documents, over twelve thousand articles from IEEE publications from the period of 1995-2002, would be provided.

2.2.2 Characteristics of Web pages

A number of researchers have explored the potential of hyperlinks in the document clustering process. R. Weiss et al. (1996) were assigning higher similarities to documents that have ancestors and descendants in common; their preliminary results illustrated that combining term and link information yields improved results. Y. Wang & M. Kitsuregawa (2002) experimented with best ways of combining terms from Web pages with words from in-link pages (pointing to the Web page) and out-link pages (leading from the Web page), and achieved improved results.

Other Web-specific characteristics have been explored. Information about users' traversals in the category structure has been experimented with (Chen, LaPaugh & Singh 2002); as well as usage logs utilized as (Su et al. 2001). The hypothesis behind this approach is that the relevancy information is objectively reflected by the usage logs; for example, it is assumed that frequent visits by the same person to two seemingly unrelated documents indicate that they are closely related.

2.2.3 Application

Clustering is the unsupervised classification of objects, based on patterns (observations, data items, feature vectors) into groups or clusters (Jain, Murty & Flynn 1999, p.264). It has been addressed in various disciplines for many different applications (ibid.); in information retrieval, documents are the ones that are grouped or clustered (hence the term *document clustering*).

Traditionally, document clustering has been applied to improve document retrieval (for a review, see Willet 1988; for an example, see Tombros & Rijsbergen 2001). In this paper the emphasis is on automated generation of hierarchical clusters structure and subsequent assignment of documents to those clusters for browsing.

An early attempt to cluster a document collection into clusters for the purpose of browsing was Scatter/Gather (Cutting et al. 1992). Scatter/Gather would partition the collection into clusters of related documents, present summaries of the clusters to the user for selection, and when the user would select a cluster, the narrower clusters were presented; when the narrowest cluster would be reached, documents were enumerated. Another approach is presented by M. Merchkour, D. Harper & G. Muresan (1998). First the so-called source collection (an authoritative collection representative in the domain of interest of the users) would be clustered for the user to browse it, with the purpose of helping him/her with defining the query. Then the query would be submitted via a Web search engine to the target collection, which is the World Wide Web. The results would be clustered into the same categories as in the source collection. H. Kim & P. Chan (2003) attempted to build a personalized hierarchy for an individual user, from a set of Web pages the user visited, by clustering words from those pages. Other research has been conducted in automated construction of vocabularies for browsing (see, for example, Chakrabarti et al. 1998; Wacholder, Evans & Klavans 2001).

Another application of automated generation of hierarchical category structure and subsequent assignment of documents to those categories is organization of Web search engine results (see, for example, Clusty the clustering engine 2004; MetaCrawler Web search 2005; Zamir et al. 1997; Zamir & Etzioni, 1998; Palmer et al. 2001; Wang & Kitsuregawa 2002).

2.2.4 Summary

Like in text categorization, documents are first represented as vectors of term weights. Then they are compared for similarity, and grouped into partitional

or hierarchical clusters using different algorithms. Characteristics of Web documents similar to those from text categorization approach have been explored.

In evaluation, precision, recall and other measures are used, while end-users and subject experts are usually left out.

Unlike text categorization, document clustering doesn't require either training documents, or pre-existing categories into which the documents are to be grouped. The categories are created when groups are formed – thus, both the names of the groups and relationships between them are automatically derived. The derivation of names and relationships is at the same time the most challenging issue in document clustering.

Document clustering was traditionally used to improve information retrieval. Today it is better suited for clustering search-engine results than for organizing a collection of documents for browsing, because automatically derived cluster labels and relationships between the clusters are incorrect or inconsistent. Also, clusters change as new documents are added to the collection – such instability of browsing structure is not user-friendly either.

2.3 Document classification

2.3.1 Special features

Description of features

Document classification is a library science approach. The tradition of automating the process of subject determination of a document and assigning it to a term from a controlled vocabulary partly has its roots in machine-aided indexing (MAI). MAI has been used to suggest controlled vocabulary terms to be assigned to a document.

The automated part of this approach differs from the previous two in that it is generally not based on either supervised or unsupervised learning. Neither do documents and classes get represented by vectors. In document classification, the algorithm typically compares extracted terms from the text to be classified, to mapped terms from the controlled vocabulary (string-to-string matching). At the same time, this approach does share similarities with text categorization and document clustering: the pre-processing of documents to be classified includes stop-words removal; stemming can be conducted; words or phrases from the text of documents to be classified are extracted and weights are assigned to them based on different heuristics; Web-page characteristics have been explored, although to a lesser degree.

The most important part of this approach is controlled vocabularies, most of which have been created and maintained for use in libraries and indexing and abstracting services, some of them for more than a century. These vocabularies have devices to "control" polysemy, synonymy, and homonymy of the natural language. They can have systematic hierarchies of concepts, and a variety of relationships defined between the concepts. There are different types of controlled vocabularies, such as classification schemes, thesauri and subject heading systems. With the World Wide Web, new types of vocabularies emerged within the computer science and the Semantic Web communities: ontologies and search-engine directories of Web pages. All these vocabularies have distinct characteristics and are consequently better suited for some classification tasks and applications than others (Koch & Day 1997; Koch and Zettergren 1999; see also Vizine-Goetz 1996). For example, subject headings systems normally do not have detailed hierarchies of terms (exception: Medical Subject Headings), while classification schemes consist of hierarchically structured groups of classes. The latter are better suited for subject browsing. Also, different classification schemes have different characteristics of hierarchical levels. For subject browsing the following are important: the bigger the collection, the more depth should the hierarchy contain; hierarchically flat schemes are not effective for browsing; classes should contain more than just one or two documents (Schwartz 2001, p.48). On the other hand, subject heading systems and thesauri have traditionally been developed for subject indexing that would describe topics of the document as specifically as possible. Since both classification schemes and subject headings or thesauri provide users with different aspects of subject information and different searching functions, their combined usage has been part of practice in indexing and abstracting services. Ontologies are usually designed for very specific subject areas and provide rich relationships between terms. Search-engine directories and other home-grown schemes on the Web, "…even those with well-developed terminological policies such as Yahoo… suffer from a lack of understanding of principles of classification design and development. The larger the collection grows, the more confusing and overwhelming a poorly designed hierarchy becomes…" (Schwartz 2001, p.76).

Although well structured and developed, existing controlled vocabularies need to be improved for the new roles in the electronic environment. Adjustments should include: 1) improved currency and capability for accommodating new terminology, 2) flexibility and expandability – including possibilities for decomposing faceted notation for retrieval purposes, 3) intelligibility, intuitiveness, and transparency – it should

be easy for the user to use, responsive to individual learning styles, able to adjust to the interests of users, and allow for custom views, 4) universality – the scheme should be applicable for different types of collections and communities and should be able to be integrated with other subject languages, 5) authoritativeness – there should be a method of reaching consensus on terminology, structure, revision, and so on, but that consensus should include user communities ([10], p.77-78). Some of the controlled vocabularies are already being adjusted, such as: AGROVOC, the agricultural thesaurus, (Soergel et al. 2004), WebDewey, which is the Dewey Decimal Classification adapted for the electronic environment (About DDC : research : a vital part of ongoing development 2005), California Environmental Resources thesaurus (CERES thesaurus effort 2003).

2.3.1.2 Differences within the approach

The differences occur in document pre-processing, which includes word or phrase extraction, stemming etc., heuristic principles (such as weighting based on where the term/word occurs), occurrence frequency, linguistic methods, and controlled vocabulary applied.

The first major project aimed at automated classification of Web pages based on a controlled vocabulary was the Nordic WAIS/World Wide Web Project (Nordic WAIS/World Wide Web Project 1995), which took place at Lund University Library and National Technological Library of Denmark (Ardö et al. 1994; Koch 1994). In this project automated classification of the World Wide Web and WAIS (Wide Area Information Server) databases using Universal Decimal Classification (UDC) was experimented with. A WAIS subject tree was built based on two top levels of UDC, i.e. 51 classes. The process involved the following steps: words from different parts of database descriptions were extracted, and they were weighted based on which part of the description they belonged to; by comparing the extracted words with UDC's vocabulary a ranked list of suggested classifications was generated. The project started in 1993, and ended in 1996, when WAIS databases came out of fashion.

GERHARD is a robot-generated Web index of Web documents in Germany (GERHARD : German Harvest Automated Retrieval and Directory 1998; Möller et al. 1999; GERHARD – Navigating the Web with the Universal Decimal Classification System 1999). It is based on a multilingual version of UDC in English, German and French, adapted by the Swiss Federal Institute of Technology Zurich (Eidgenössische Technische Hochschule Zürich – ETHZ). GERHARD's approach included advanced linguistic analysis: from captions, stop words were removed, each word was

morphologically analysed and reduced to stem; from Web pages stop words were also removed and prefixes were cut off. After the linguistic analysis, phrases were extracted from the Web pages and matched against the captions. The resulting set of UDC notations was ranked and weighted statistically, according to frequencies and document structure.

Online Computer Library Center's (OCLC) project Scorpion (Scorpion 2004) built tools for automated subject recognition, using Dewey Decimal Classification (DDC). The main idea was to treat a document to be indexed as a query against the DDC knowledge base. The results of the "search" were treated as subjects of the document. R. Larson (1992) used this idea earlier, for books. In Scorpion clustering was also used, for refining the result set and for further grouping of documents falling in the same DDC class (Subramanian, Shafer 1998). The SMART (System for Manipulating And Retrieving Text) weighting scheme was used, in which term weights were calculated based on several parameters: the number of times that the term occurred in a record; how important the term was to the entire collection based on the number of records in which it occurred; and, the normalization value, which is the cosine normalization that computes the angle between vector representations of a record and a query. Different combinations of these elements have been experimented with. Another OCLC project, WordSmith (Godby & Reighart 1998), was to develop software to extract significant noun phrases from a document. The idea behind it was that the precision of automated classification could be improved if the input to the classifier were represented as a list of the most significant noun phrases, instead as the complete text of the raw document. However, it showed that there were no significant differences. OCLC currently works on releasing FAST (FAST : Faceted Application of Subject Terminology 2005), based on the Library of Congress Subject Headings (LCSH), which are modified into a post-coordinated faceted vocabulary. The eight facets to be implemented are: Topical, Geographic (Place), Personal Name, Corporate Name, Form (Type, Genre), Chronological (Time, Period), Title and Meeting Place. FAST could also serve as a knowledge base for automated classification, like the DDC database did in Scorpion (FAST as a knowledge base for automatic classification 2003).

WWLib (Wolverhampton Web Library) is a manually maintained library catalogue of British Web resources, within which experiments on automating its processes were conducted (Wallis & Burden 1995; Jenkins et al. 1998). Original classifier from 1995 was based on comparing text from each document to DDC captions. In 1998 each classmark in the DDC captions file was enriched with additional keywords and synonyms. Keywords

extracted from the document were weighted on the basis of their position in the document. The classifier began by matching documents against class representatives of top ten DDC classes and then proceeded down through the hierarchy to those subclasses that had a significant measure of similarity (Dice's coefficient) with the document.

"All" Engineering ("All" Engineering resources on the Internet : a companion service to EELS 2003) is a robot-generated Web index of about 300000 Web documents, developed within DESIRE (DESIRE project 1999; DESIRE : Development of a European Service for Information on Research and Education 2000), as an experimental module of the manually created subject gateway Engineering Electronic Library (EELS) (Koch & Ardö 2000; Engineering Electronic Library 2003). Engineering Index (Ei) thesaurus was used; in this thesaurus, terms are enriched with their mappings to Ei classes. Both Ei captions and thesaurus terms were matched against the extracted title, metadata, headings and plain text of a full-text document from the World Wide Web. Weighting was based on term complexity and type of classification, location and frequency. Each pair of term-class codes was assigned a weight depending on the type of term (Boolean, phrase, single word), and the type of class code (main code, the class to be used for the term, or optional code, the class to be used under certain circumstances); a match of a Boolean expression or a phrase was made more discriminating than a match of a single word; a main code was made more important than an optional code. Having experimented with different approaches for stemming and stop-word removal, the best results were gained when an expanded stop-word list was used, and stemming was not applied. The DESIRE project proved the importance of applying a good controlled vocabulary in achieving the classification accuracy: 60% of documents were correctly classified, using only a very simple algorithm based on a limited set of heuristics and simple weighting. Another robot-generated Web index, Engine-e (Engine-e 2004), used a slightly modified automated classification approach to the one developed in "All" Engineering (Lindholm, Schönthal & Jansson 2003). Engine-e provided subject browsing of engineering documents based on Ei terms, with six broader categories as starting points.

The project BINDEX (Bilingual Automatic Parallel Indexing and Classification) (HLT Project Factsheet : BINDEX 2001; Nübel at al. 2002) was aimed at indexing and classifying abstracts from engineering in English and German, using English INSPEC thesaurus and INSPEC classification, FIZ Technik's bilingual Thesaurus "Engineering and Management" and the Classification Scheme "Fachordnung Technik 1997". They performed morpho-syntactic analysis of a document, which consisted of identification of

single and multiple-word terms, tagging and lemmatization, and homograph resolution. The extracted keywords were checked against the INSPEC thesaurus and the German part of "Engineering and Management", and classification codes were derived. Keywords which were not in the thesaurus, were assigned as free indexing terms.

2.3.1.3 Evaluation methods

Measures such as precision and recall have been used. This approach differs from the other two approaches in that evaluation of document classification tends to also involve subject experts or intended users (see, for example, Koch & Ardö 2000), which is in line with traditional library science evaluations.

Examples of data collections that have been used are harvested Web documents (GERHARD, "All" Engineering), and bibliographic records of Internet resources (Scorpion).

2.3.2 Summary

Document classification is a library science approach. It differs from text categorization and document clustering in that well-developed controlled vocabularies are employed, whereas vector space model and algorithms based on vector calculations are generally not used. Instead, selected terms from documents to be classified are compared against terms in the chosen controlled vocabulary, whereby often computational linguistic techniques are employed.

In evaluation, performance measures from information retrieval are used, and, unlike in the other two approaches, subject experts or users tend to be involved.

In the focus of research are mainly (publicly available) operative information systems that provide browsing access to their document collections.

2.4 Mixed approach

Mixed approach is the term used here to refer to a machine-learning or an information-retrieval approach, in which also controlled vocabularies that have been traditionally used in libraries and indexing and abstracting services are used. There do not seem to be many examples of this approach. E. Frank & G. Paynter (2004) applied machine-learning techniques to assign LCC

notations to resources that already have an LCSH term assigned. Their solution has been applied to INFOMINE (subject gateway for scholarly resources, http://infomine.ucr.edu/), where it is used to support hierarchical browsing. There are also cases in which search engine results were grouped into pre-existing subject categories for browsing. For example, W. Pratt (1997) who experimented with organizing search results into MeSH categories.

Other mixed approaches are also possible, such as the one applied in the Scorpion project (see section 2.3.1.2).

The emergence of this approach demonstrates the potentials for utilizing ideas and methods from another community's approach.

3 Discussion

3.1 Features of automated classification approaches

Several problems with automated classification in general have been identified in the literature. As E. Svenonius (2000, p.46-49) claims, automating subject determination belongs to logical positivism – a subject is considered to be a string occurring above a certain frequency, is not a stop word and is in a given location, such as a title. Algorithms in all the approaches are based on statistical, locative or linguistic data, or, like in clustering algorithms, inferences are made such as "if document A is on subject X, then if document B is sufficiently similar to document A (above a certain threshold), then document B is on that subject." It is assumed that concepts have names, which is common in science, but is not always the case in humanities and social sciences. Automated classification in certain domains has been entirely unexplored, due to lack of suitable data collections or good-quality controlled vocabularies. Another critique given is the lack of theoretical justifications for vector manipulations, such as the cosine measure that is used to obtain vector similarities (Salton 1991, p.975).

In regards to similarities and differences between the approaches, document pre-processing (e.g. selection of terms) is common to all the approaches. Various Web page characteristics have also been explored by all the three communities, although mostly within the text categorization approach. Major differences between the three approaches are in applied algorithms, employment or not of the vector-space model and of controlled vocabularies, especially as to how well suited they are for subject browsing (cf. 3.3 Application for subject browsing). Since there are similarities between

approaches, the hypothesis is that idea exchange and co-operation between the three communities would be beneficial. The hypothesis does seem to be supported by the emergence of the mixed approach. They could all benefit from at least looking into each other's approaches to document pre-processing and indexing, and exchanging ideas about properties of Web pages and how they could be used. However, there seems to be little co-operation or idea exchange among them. This is also supported by the fact that, to the author's knowledge, no review paper on automated classification attempted to discuss more than one community's approach. A recent bibliometric study (Golub & Larsen 2005) shows that the three communities are quite clearly mutually independent when looking at citation patterns; and that document clustering and text categorization are closer to each other, while the document classification community is almost entirely isolated. Further research is needed to determine why direct and indirect links are lacking between the document classification and the other two communities, in spite of emergence of the mixed approach.

3.2 Evaluation

The problem of deriving the correct interpretation of a document's subject matter has been much discussed in the library science literature (while much less so in machine learning and information retrieval communities). It has been reported that different people, whether users or subject indexers, would assign different subject terms or classes to the same document. Studies on inter-indexer and intra-indexer consistency report generally low indexer consistency (Olson & Boll, p.99-101). There are two main factors that seem to affect it: 1) higher specificity and higher exhaustivity both lead to lower consistency (indexers choose the same first term for the major subject of the document, but the consistency will decrease as they choose more terms); 2) the bigger the vocabulary, or, the more choices the indexers have, the less likely they will choose the same terms (ibid.). Apart from exhaustivity or specificity of subject indexing and vocabulary size, the purpose that the document collection is to serve is another important factor in deciding which classes or terms are to be chosen or made more prominent.

Having the above in mind, performance measures need to be questioned and evaluation has to be dealt with in the broader contexts of users and their tasks. Subject experts or intended end-users have been mostly excluded from evaluation in text categorization and document clustering approaches, while document classification approach tends to involve them to

a larger degree, which corresponds to the tradition of evaluating library services in general.

Due to poor evaluation, it is difficult to estimate to what degree the automated classification tools of today are really applicable in operative information systems and for which tasks.

3.3 Application for subject browsing

Research in text categorization seems to be mainly in improving categorization performance, and experiments are conducted under controlled conditions. Research in which Web pages have been categorized into hierarchical structures for browsing generally does not involve well-developed classification schemes, but home-grown structures such as directories of search engines that are not structured and maintained well enough.

In document clustering, clusters' labels and relationships between the clusters are automatically produced. Labelling of the clusters is a major research problem, with relationships between the categories, such as those of equivalence, related-term and hierarchical relationships, being even more difficult to automatically derive (Svenonius 2000, p.168). "Automatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand" (Chen & Dumais 2000). Also, clusters change as new documents are added to the collection. Unstable category names in Web services and digital libraries, for example, are not user-friendly. T. Koch & A. Zettergren (1999) suggest that document clustering is better suited for organizing Web search engine results.

Document classification approach employs well-developed classification schemes, which are suitable for subject browsing. However, the future research should include improving controlled vocabularies for browsing in the electronic environment, as well as making them more suitable for automated classification.

Acknowledgements

Many thanks to Traugott Koch, Anders Ardö, Tatjana Aparac Jelušić, Johan Eklund, Ingo Frommholz, Repke de Vries and the Journal of Documentation reviewers for providing valuable feedback on the paper.

Bibliography

1. " "All" Engineering resources on the Internet : a companion service to EELS", (31 January 2003), (*EELS, Engineering E-Library, Sweden*), Available: http://eels.lub.lu.se/ae/ (Accessed: 22 December 2004).

2. "20 Newsgroups DataSet", (January 1998), (*The 4 Universities Data Set*), Available: http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html (Accessed: 22 December 2004).

3. "About DDC : research : a vital part of ongoing development", (2005) (*Dewey Services*), Available: http://www.oclc.org/dewey/about/research/ (Accessed: 8 August 2005).

4. Ardö, A. et al. (1994), "Improving resource discovery and retrieval on the Internet : The Nordic WAIS/World Wide Web project summary report", *NORDINFO Nytt*, vol. 17, no. 4, pp. 13-28.

5. Attardi, G., Gullì, A., and Sebastiani, F. (1999), 'Automatic Web Page Categorization by Link and Context Analysis' In: Hutchison, C., Lanzarone, G. (eds), *Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pp. 105-119.

6. Bekkerman, R. et al. (2003), "Distributional Word Clusters vs. Words for Text Categorization", *Journal of Machine Learning Research*, vol. 3, pp. 1183-1208.

7. Blum, A., and Mitchell, T. (1998), 'Combining labeled and unlabeled data with co-training', In: *COLT: Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann Publishers.

8. Cai, L., and Hofmann, T. (2003), 'Text categorization by boosting automatically extracted concepts', In: Callan, J. et al. (eds), *Proceedings of SIGIR-03, 26th ACM International Conference on Research and Development in Information Retrieval,* pp. 182—189.

9. "CERES thesaurus effort". (22 December 2003), (*CERES The California Environmental Resources Evaluation System*), Available: http://ceres.ca.gov/thesaurus/ (Accessed: 22 December 2004).

10. Chakrabarti, S. et al. (1998), 'Automatic resource compilation by analyzing hyperlink structure and associated text', In: *Proceedings of the seventh international conference on World Wide Web 7*, Brisbane, Australia, pp. 65 – 74.

11. Chakrabarti, S., Dom, B., and Indyk, P. (1998), "Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies", *Journal of Very Large Data Bases*, 1998, vol. 7, no. 3, pp. 163-178.

12. Chan, L.M. (1994), *Cataloging and Classification : An Introduction*, 2nd ed., McGraw-Hill, New York.

13. Chen, H., and Dumais, S.T. (2000), 'Bringing Order to the Web: Automatically Categorizing Search Results', In: *Proceedings of CHI-00, ACM International Conference on Human Factors in Computing Systems, Den Haag*, pp. 145–152.

14. Chen, M., LaPaugh, A., and Singh, J.P. (2002), 'Categorizing information objects from user access patterns', In: *Proceedings of the eleventh international conference on Information and knowledge management*, November 04-09, pp. 365–372.

15. "Clusty the clustering engine", (2004), (*Vivisimo*), Available: http://www.clusty.com (Accessed: 22 December 2004).

16. Cutting, D. et al. (1992), 'Scatter/Gather : A Cluster-based Approach to Browsing Large Document Collections', In: *Proceedings of the 15th Annual International ACM/SIGIR Conference, Copenhagen,* pp. 318-329.

17. "DESIRE : Development of a European Service for Information on Research and Education" (07 August 2000) (*DESIRE*), Available: http://www.desire.org/ (Accessed: 22 December 2004).

18. "DESIRE project". (30 March 1999) (*Lunds Universitets Bibliotek*). Available: http://www.lub.lu.se/desire (Accessed: 22 December 2004).

19. Dittenbach, M., Berger, H., and Merkl, D. (2004), 'Improving domain ontologies by mining semantics from text', In: *Proceedings of the first Asian-Pacific conference on Conceptual modeling*, Dunedin, New Zealand, vol. 31, pp. 91-100.

20. Dumais, S.T., and Chen, H. (2000), 'Hierarchical classification of web content', In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, July 24-28, 2000, Athens, Greece, pp. 256-263.

21. Dumais, S.T., Lewis, D.D., and Sebastiani, F. (2002), 'Report on the Workshop on Operational Text Classification Systems (OTC-02)', In: *ACM SIGIR Forum,* vol. 35, no. 2, 2001, pp. 8-11.

22. "Engine-e", (13 February 2004), (*Lund University Libraries*), Available: http://engine-e.lub.lu.se/ (Accessed: 22 December 2004).

23. "Engineering Electronic Library", (31 January 2003) (*Lund University Libraries*), Available: http://eels.lub.lu.se/ (Accessed: 22 December 2004).

24. "FAST as a knowledge base for automated classification", (2003), (*OCLC projects*), Available: http://www.oclc.org/research/projects/fastac/ (Accessed: 7 August 2005).

25. "FAST : Faceted Application of Subject Terminology", (*OCLC projects*), Available: http://www.oclc.org/research/projects/fast/ (Accessed: 22 December 2004).

26. Fasulo, D. (1999), 'An analysis of recent work on clustering algorithms : technical report', University of Washington, 1999. Available: http://citeseer.nj.nec.com/fasulo99analysi.html (Accessed: 22 December 2004).

27. Fisher, M., and Everson R. (2003), 'When are links useful? Experiments in text classification', In: *Proceedings of ECIR-03, 25th European Conference on Information Retrieval*, Pisa, IT, pp. 41-56.

28. Frank, E., and Paynter, G.W. (2004), "Predicting Library of Congress Classifications From Library of Congress Subject Headings", *Journal of the American Society for Information Science and Technology*, vol. 55, no. 3, pp. 214–227.

29. Fürnkranz, J. (1999), 'Exploiting Structural Information for Text Classification on the WWW', In: *Proceedings of IDA-99, 3rd Symposium on Intelligent Data Analysis*, pp. 487-497.

30. Fürnkranz, J. (2002), "Hyperlink Ensembles : A Case Study in Hypertext Classification", *Information Fusion*, vol. 3, no. 4, pp. 299-312.

31. Garfield, E., Malin, M.V., and Small, H. (1975), "A System for Automatic Classification of Scientific Literature", Reprinted from *Journal of the Indian Institute of Science* vol. 57, no. 2, pp. 61-74. (Reprinted in: Essays of an Information Scientist, vol. 2, pp. 356-365).

32. "GERHARD - Navigating the Web with the Universal Decimal Classification System" (September 1999), (*GERHARD*), Available: http://www.gerhard.de/info/dokumente/vortraege/ecdl99/html/index.htm (Accessed: 22 December 2004).

33. "GERHARD : German Harvest Automated Retrieval and Directory", (20 July 1998), (*GERHARD*), Available: http://www.gerhard.de/ (Accessed: 22 December 2004).

34. Ghani, R., Slattery, S., and Yang, Y. (2001), 'Hypertext Categorization using Hyperlink Patterns and Metadata', In: *Proceedings of ICML-01, 18th International Conference on Machine Learning*, pp. 178-185.

35. Glover, E.J. et al. (2002), 'Using Web structure for Classifying and Describing Web Pages', In: *Proceedings of the eleventh international conference on World Wide Web* Honolulu, Hawaii, USA, pp. 562–569.

36. Glover, E.J. et al. (2003), 'Inferring Hierarchical Descriptions', In: *Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM 2002*, November 4-9, 2002, pp. 507-514.

37. Godby, J., and Reighart, R. "The WordSmith indexing system," (1998), (*OCLC Digital Archive*), Available: http://digitalarchive.oclc.org/da/ViewObject.jsp?fileid=0000003487:000000090408&reqid=33836 (Accessed: 22 December 2004).

38. Golub, K. and Larsen, B. (2005), 'Different Approaches to Automated Classification: Is There an Exchange of Ideas?' In: Ingwersen, P. and Larsen, B. eds. *Proceedings of ISSI 2005 - the 10th International Conference of the International Society for Scientometrics and Informetrics,* Stockholm, Sweden, July 24-28, 2005, Volume 1. Stockholm: Karolinska University Press, pp. 270-274.

39. Goren-Bar, D. et al. (2000), 'Supervised Learning for Automatic Classification of Documents using Self-Organizing Maps', *Proceedings of the First DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries*, Zűrich, Switzerland, 11-12 December, 2000.

40. Gövert, N., Lalmas, M., and Fuhr, N. (1999), 'A probabilistic description-oriented approach for categorising Web documents', In: Proceedings of the Eighth International Conference on Information and Knowledge Management, pp. 475-482.

41. Hartigan, J.A. (1996), Introduction, in *Clustering and classification* Arabie, P., Hubert, L., De Soete, G. (eds), World Scientific, Singapore.

42. Hatzivassiloglou, V., Gravano, L., and Maganti, A. (2000), 'An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering', *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval,* Athens, Greece, pp. 224-231.

43. Haveliwala, T.H., Gionis, A., and Indyk, P. (2000), 'Scalable techniques for clustering the Web', In: *Third International Workshop on the Web and Databases, May 2000*, pp. 129-134.

44. Heuser, U., Babanine, A., and Rosenstiel, W. (1998), 'HTML Documents Classification using (Non-linear) Principal Component Analysis and Self-Organizing Maps', In: *Proc. of the Fourth International Conference on Neural Networks and their Applications (Neurap'98)*, March 11-13, 1998, Marseilles, France, pp. 291-295

45. Hersh, W.R. (1994), 'OHSUMED : An interactive retrieval evaluation and new large test collection for research', In: *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 192-201.

46. "HLT Project Factsheet : BINDEX", (14 November 2001), (*HLTCentral*), Available: http://www.hltcentral.org/projects/print.php?acronym=BINDEX (Accessed: 22 December 2004).

47. INitiative for the Evaluation of XML Retrieval (INEX), (2 December 2004) (*DELOS Network of Excellence for Digital Libraries*), Available: http://inex.is.informatik.uni-duisburg.de/ (Accessed: 22 December 2004).

48. Jain, A.K., Murty, M.N., and Flynn, P.J. (1999), "Data clustering : a review", *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323.

49. Jenkins, C. et al. (1998), "Automatic Classification of Web Resources using Java and Dewey Decimal Classification", *Computer Networks & Isdn Systems*, vol. 30, pp. 646-648.

50. Kim, H.R., and Chan, P.K. (2003), 'Learning Implicit User Interest Hierarchy for Context in Personalization', In: *Proc. Intl. Conf. on Intelligent User Interfaces*, pp. 101-108.

51. Koch, T. (1994), 'Experiments with Automatic Classification of WAIS Databases and Indexing of WWW', In *Internet World & Document Delivery World International 94*, London, May 1994, pp. 112-115.

52. Koch, T., and Ardö, A. (2000), "Automatic classification", (11 February 2000), (*DESIRE II D3.6a, Overview of results),* Available: http://www.lub.lu.se/desire/DESIRE36a-overview.html (Accessed: 22 December 2004).

53. Koch, T., and Day, M. (1997), "The role of classification schemes in Internet resource description and discovery" *(EU Project DESIRE, Deliverable D3.2.3)*, Available: http://www.lub.lu.se/desire/radar/reports/D3.2.3/ (Accessed: 22 December 2004).

54. Koch, T., and Zettergren, A.-S. (1999), Provide browsing in subject gateways using classification schemes, (30 March 1999), (*EU Project DESIRE II*) http://www.lub.lu.se/desire/handbook/class.html (Accessed: 22 December 2004).

55. Kohonen, T. (2001), *Self-Organizing Maps*, 3rd ed., Springer-Verlag, Berlin.

56. Koller, D., and Sahami, M. (1997), 'Hierarchically classifying documents using very few words', In: *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pp. 170-178.

57. Labrou, Y., and Finin, T. (1999), 'Yahoo! as an ontology : using Yahoo! categories to describe documents', In: *Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management*, pp. 180-187.

58. Larson, R.R. (1992), "Experiments in automatic Library of Congress Classification", *Journal of the American Society for Information Science*, vol. 43, no. 2, pp. 130-148.

59. Li, Y.H., and Jain, A.K. (1998), "Classification of text documents", *The Computer Journal*, vol. 41, no. 8, pp. 537-546.

60. Liere, R., and Tadepalli, P. (1998), 'Active Learning with Committees : Preliminary Results in Comparing Winnow and Perceptron in Text Categorization', *Proceedings of CONALD-98, 1st Conference on Automated Learning and Discovery.*

61. Lindholm, J., Schönthal, T., and Jansson, K. (2003), "Experiences of Harvesting Web Resources in Engineering using Automatic Classification", *Ariadne*, no. 37. Available at: http://www.ariadne.ac.uk/issue37/lindholm/.

62. Liu, X. et al. (2002), 'Document Clustering with Cluster Refinement and Model Selection Capabilities', In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval,* Tampere, Finland, pp. 191-198.

63. Mandhani, B., Joshi , S. and Kummamuru K. (2003), 'A matrix density based algorithm to hierarchically co-cluster documents and words', In: *Proceedings of the twelfth international conference on World Wide Web*, Budapest, Hungary, pp. 511-518.

64. Manning, C. and Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge.

65. Merchkour, M., Harper, D.J. and Muresan, G. (1998), 'The WebCluster project : Using clustering for mediating access to the World Wide Web' In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne, Australia, pp. 357-358.

66. McCallum, A. et al. (1998), 'Improving text classification by shrinkage in a hierarchy of classes', In: *ICML-98, 15th International Conference on Machine Learning,* pp. 359-367.

67. McCallum et al. (1999), 'Building Domain-Specific Search Engines with Machine Learning Techniques', In: *AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace.*

68. McCallum, A. et al. (2000), "Automating the Construction of Internet Portals with Machine Learning", *Information Retrieval Journal*, vol. 3, pp. 127-163.

69. "MetaCrawler Web Search", (2005), Available: http://metacrawler.com (Accessed: 5 August 2005).

70. Mitchell, T. (1997), *Machine Learning.* McGraw Hill, New York.

71. Mladenic, D. (1998), 'Turning Yahoo into an Automatic Web-Page Classifier', In: *Proceedings of the 13th European Conference on Artificial Intelligence ECAI'98,* pp. 473-474.

72. Mladenic, D. and Grobelnik, M. (2003), "Feature selection on hierarchy of Web documents", *Decision Support Systems*, vol. 35, no. 1, pp. 45-87.

73. Möller, G. et al. (1999), 'Automatic classification of the WWW using the Universal Decimal Classification', In: McKenna, B. (ed), *Proceedings of the 23rd International Online Information Meeting. London, 7-9 Dec 1999*, pp. 231-238.

74. Nordic WAIS/World Wide Web Project, (14 February 1995), (*Lund University Libraries*), Available: http://www.lub.lu.se/W4/ (Accessed: 22 December 2004).

75. Nübel, R. et al. "Bilingual indexing for information retrieval with AUTINDEX". In: *LREC Proceedings,* Las Palmas 2002.

76. Olson, H.A., and Boll, J.J. (2001), *Subject analysis in online catalogs*, 2nd ed, Libraries Unlimited, Englewood, Colorado.

77. Poincot, P., Lesteven, P.S., and Murtagh, F. (1998), "A spatial user interface to the astronomical literature", *Astronomy & Astrophysics*, May II 1998, pp. 183-191.

78. Palmer, C.R. et al. (2001), 'Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results', In: *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries,* Roanoke, Virginia, 2001, pp. 451.

79. Pierre, J.M. (2001), "On the automated classification of Web sites", *Linköping Electronic Articles in Computer and Information Science*, vol. 6, no. 001.

80. Pratt, W. (1997), 'Dynamic organization of search results using the UMLS', In: *American Medical Informatics Association Fall Symposium*, 1997, pp. 480-484.

81. Rasmussen, E. (1992), Clustering algoritms, in Frakes, W.B., Baeza-Yates, R. (eds), *Information retrieval : data structures and algorithms*, Prentice Hall, Engelwood Cliffs.

82. Rauber, A., Merkl, D. (1999), 'SOMLib : A Digital Library System Based on Neural Networks', In: *Proceedings of the fourth ACM conference on Digital libraries*, Berkeley, California, United States, pp. 240-241.

83. "Reuters-21578", (2004), Available: http://www.daviddlewis.com/resources/testcollections/reuters21578/, (Accessed: 3 August 2005).

84. Rocchio, J.J. (1971), 'Relevance feedback in information retrieval', In: Salton, G. (ed) *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, NJ, pp. 313–323.

85. Ruiz, M.E. and Srinivasan, P. (1999), 'Hierarchical neural networks for text categorization', In: *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 281-282.

86. Sahami, M., Yusufali, M., and Baldonado, M.Q. (1998), 'SONIA: a service for organizing networked information autonomously', In: *3rd ACM conference on digital libraries*, Pittsburgh, pp. 200-209.

87. Salton, G. (1991), "Developments in automatic text retrieval", *Science*, vol. 253, pp. 974-979.

88. Schweighofer, E., Rauber, A. and Dittenbach, M. (2001), 'Automatic text representation, classification and labeling in European law', In: *ICAIL 2001*, pp. 78-87.

89. Schütze, H., Hull, D.A., and Pedersen, J.O. (1995), 'A comparison of classifiers and document representations for the routing problem', In: *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, pp. 229-237.

90. Schwartz, C. (2001), *Sorting out the web : Approaches to subject access,* Ablex, Westport, CT.

91. "Scorpion", (2004), (*OCLC software*) Available: http://www.oclc.org/research/software/scorpion/default.htm (Accessed: 22 December 2004).

92. Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47.

93. Slattery, S., and Craven, M. (2000), 'Discovering test set regularities in relational domains', In: *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pp. 895-902.

94. Slonim, N., Friedman, N., and Tishby, N. (2003), 'Unsupervised Document Classification using Sequential Information Maximization', In: *Proceedings of SIGIR'02, 25th ACM International Conference on Research and Development of Information Retireval*, Tampere, Finland, 2002.

95. Soergel, D. et al. (2004), "Reengineering Thesauri for New Applications : The AGROVOC Example", *Journal of Digital Information*, vol. 4, no. 4, article no. 257, http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/.

96. Steinbach, M., Karypis, G., and Kumar, V. (2000), 'A comparison of document clustering techniques', In: *KDD Workshop on Text Mining,* Boston, MA, August 20-23, 2000.

97. Su, Z. et al. (2001), 'Correlation-based Document Clustering using Web Logs', In: *Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)*, January 03-06, 2001, vol. 5, pp. 5022.

98. Subramanian, S., and Shafer, K.E. "Clustering", (1998), (*OCLC Publications*), Available: http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003409, (Accessed: 22 December 2004).

99. Sun, A., Lim, E.-P., and Ng, W.-K. (2001), 'Hierarchical Text Classification and Evaluation', In *ICDM 2001, IEEE Int. Conf. on Data Mining.*

100. Svenonius, E. (2000), *The intellectual foundations of information organization*. MIT Press, Cambridge, MA.

101. "Thunderstone's Web Site Catalog", (2005), Available: http://search.thunderstone.com/texis/websearch (Accessed: 4 August 2005).

102. Tombros, A., and van Rijsbergen, C.J. (2001), 'Query-Sensitive Similarity Measures for the Calculation of Interdocument Relationships', In: *Proceedings of the tenth international conference on Information and knowledge management,* Atlanta, Georgia, USA, pp. 17-24.

103. Toth E. (2002), "Innovative solutions in automatic classification : A brief summary", *Libri*, vol. 25, no. 1, pp. 48-53.

104. "TREC : Text REtrieval Conference", (15 December 2004), (*National Institute of Standards and Technology*), Available: http://trec.nist.gov/ (Accessed: 22 December 2004).

105. Wacholder, N., Evans, D.K., and Klavans, J.L. (2001), 'Automatic Identification and Organization of Index Terms for Interactive Browsing', In: *Proceedings of the ACM-IEEE Joint Conference on Digital Libraries*, Roanoke, Virginia, June 2001, pp. 128-134.

106. Wallis, J., and Burden, P. (1995), "Towards a Classification-based Approach to Resource Discovery on the Web", (1995) (*University of Wolverhampton*), http://www.scit.wlv.ac.uk/wwlib/position.html (Accessed: 22 December 2004).

107. Wang, Y., and Kitsuregawa, M. (2002), 'Evaluating Contents-Link Coupled Web Page Clustering for Web Search Results', In: *Proceedings of the eleventh international conference on Information and knowledge management,* McLean, Virginia, USA, pp. 499-506.

108. "WebKB", (January 2001), (*CMU World Wide Knowledge Base*), Available: http://www-2.cs.cmu.edu/~webkb/, (Accessed: 22 December 2004).

109. Weiss, R., et al. (1996), 'HyPursuit : A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering', *Proceedings of the Seventh ACM Conference on Hypertext*, Washington, DC, March 1996, p. 180-193.

110. "Yahoo! Directory 2005", (2005), Available: http://dir.yahoo.com/, (Accessed: 8 August 2005).

111. Yang, C., Chen H., and Hong, K. (2003), "Visualization of large category map for Internet browsing," *Decision Support Systems (DSS)*, vol. 35, no. 1, pp. 89-102.

112. Yang, Y., Slattery, S., and Ghani, R. (2002), "A Study of Approaches to Hypertext Categorization", *Journal of Intelligent Information Systems*, vol. 8, nr. 2-3, pp. 219-241.

113. Yang, Y. (1999), "An evaluation of statistical approaches to text categorization", *Journal of Information Retrieval*, vol. 1, no. 1/2, pp. 67-88.

114. Zamir, O., and Etzioni, O. (1998), 'Web document clustering : A feasibility demonstration', In: *ACM SIGIR'98*, Australia, pp. 46-54.

115. Zamir, O. et al. (1997), 'Fast and intuitive clustering of Web documents', In: *Proceedings of the 3rd international conference on knowledge discovery and data mining*, pp. 287-290.

116. Zhao, Y., and Karypis, G. (2002), 'Evaluation of Hierarchical Clustering Algorithms for Document Dataset', In: *Proceedings of the eleventh international conference on Information and knowledge management*, McLean, Virginia, pp. 515-524.

II

Different Approaches to Automated Classification: Is There an Exchange of Ideas?

Abstract. Automated classification of text has been studied by three major research communities, machine learning, information retrieval, and library science, each taking a different approach. The paper aims to study to what a degree the three communities explore others' ideas, methods, findings. To that purpose we studied direct links (do authors from one community cite authors from another) and indirect links (using bibliographic coupling). Although the study is based on a small sample of 148 papers, the results indicate that the three communities do not exchange ideas to a great extent.

1 Introduction

Automated subject classification has been a challenging research issue for several decades. The interest has grown rapidly with the emergence of the World Wide Web (WWW) and related digital information services with very large amounts of documents, where the high costs of manual subject classification is a major hindrance.

Currently, there are three distinguishable approaches to automated subject classification of text, each taken by a different research community: text categorization, document clustering and document classification. They

differ in a number of aspects, such as: scientific tradition, methodology (including document pre-processing and indexing, test collections, characteristics of categories, evaluation methods) and application. However, all of them deal with the same problem and similarities between them exist; for example, selection of most relevant terms during document pre-processing is common to all the approaches, as is utilization of specific document characteristics. This leads one to assume that idea exchange and co-operation between the three communities would be beneficial.

The goal of the study is to examine whether simple bibliometric methods can be used to investigate to what degree the three communities utilize others' ideas, methods, and findings. Our main hypothesis is that there is hardly any exchange of ideas etc. To that purpose we studied direct links (do authors from one community cite authors from another) and indirect links using bibliographic coupling (Kessler, 1963). A freely available, offline tool for bibliometric analysis, Bibexcel, was used for the informetric analysis and map generation[1].

This paper is laid out as follows: brief descriptions of the three approaches are given in the next section, followed by a description of the methodology; results are discussed and conclusions are given in the last two sections.

2 Descriptions of the approaches

Text categorization is a machine-learning approach, in which also information retrieval methods are applied. It involves manually categorizing a number of documents to pre-defined categories (which normally lack devices for the control of polysemy, synonymy and homonymy). By learning the characteristics of those documents the automated categorization of new documents takes place. Text categorization is known as supervised learning, since the process is "supervised" by learning categories' characteristics from manually categorized documents.

Document clustering is an information-retrieval approach. Unlike text categorization, it does not involve pre-defined categories or training documents and is thus called unsupervised. The clusters and, to a limited degree, relationships between clusters are derived automatically from the documents, and the documents are subsequently assigned to those clusters.

[1] Bibexcel is developed by prof. dr. sc. Olle Persson and may be dowloaded from http://www.umu.se/inforsk.

Document classification in this paper stands for a library science approach. It involves manually created controlled vocabulary (such as classification schemes, thesauri, or subject headings systems) into categories of which documents are classified. Controlled vocabularies have devices to control the problems of polysemy, synonymy and homonymy of natural language. They have been developed and used in libraries and in indexing and abstracting services, some since the end of the 19th century.

3 Methodology

3.1 Sample

The sample consists of 148 papers related to automated classification of Web-based text resources. The majority of papers are published after 1997. Out of these 63 papers are from the information retrieval (IR) community, 52 from machine learning (ML) and 33 from the library science (LS) community. The library science set of papers include two subgroups, one 'pure' library science subgroup, and the other with papers using either IR or ML approach, but also applying controlled vocabularies such as those used by the LS community.

The sample was collected from commercially and non-commercially available databases, mostly from ACM Digital Library, ISI Web of Science as well as Web sites of projects and personal Web sites. The databases were searched for documents on automated classification of text, using a variety of search terms. Not having any formal criteria, e.g. distinct channels of publication for each community, every paper had to be at least partially read in order to be assigned to the corresponding community. Additionally, due to overlaps in content, a number of papers were assigned to two or even three categories ('mixed' category in Table 1). For more than half of the papers, records with references had to be created from scratch or converted semi-automatically. The relatively small size of the sample is due to the fact that the number of LS papers is rather small (although there are many ML and IR papers).

3.2 Informetric methods used

Two main informetric methods were chosen for the study: direct and indirect links. *Direct links* were used to determine to what extent authors from one

community cite authors from the other two communities. References of papers belonging to one community were searched for author names belonging to the other two communities. Every appearance of a name in the references was counted, which included different papers and even several instances where the searched author was cited as an editor of, e.g., conference proceedings. Only authors that were cited at least three times were examined. Authors, and not papers, were chosen because of the relatively small sample size. *Indirect links* were studied using bibliographic coupling between papers (Kessler, 1963). Bibliographic coupling was chosen because it shows the domain as it is interpreted by the researchers writing the new knowledge, and it is their own interpretation of their position in the scientific domain.

4 Results and discussion

4.1 Direct links between authors

Table 1 gives the number of times an author from one community was cited by one the other two communities, with percentages in the parentheses. They indicate that all communities cited each other. The IR and ML communities were mostly cited, and the LS community least cited.

Table 1. Number of citings between communities

	Authors cited by IR, excluding IR	**Authors cited by LS, excluding LS**	**Authors cited by ML, excluding ML**
IR	/	18 (24 %)	78 (42 %)
LS	7 (7 %)	/	29 (15 %)
ML	40 (41 %)	34 (45 %)	/
mixed	50 (52 %)	23 (31 %)	81 (43 %)

Qualitative analysis was used to determine the context in which other authors were cited. When the same paper from one community was cited by both other communities, it tended to be cited for similar reasons: either to provide an example of different classification methods and applications and compare with their own ones, or to refer to the same basic concepts of information retrieval and automatic text processing. Many of the authors citing other community's authors, also themselves belong to that community. The ML community uses IR methods and both tended to cite each other to a certain

extent. LS authors cited by the other two communities did occur, but they were restricted to the 'non-pure' LS authors and papers. There was not one single case where 'pure' LS authors were cited by either of the two other communities, and vice versa: LS authors who cited the other two communities were either 'non-pure' or belonged to another community.

4.2 Indirect links between papers

Papers, as well as references, were identified by author and labelled with the author's name, her community's tag (IR, LS, ML) and publication year. Matrixes were produced in Bibexcel and imported into a multidimensional scaling (MDS) program for creation of two-dimensional maps. The stress of scaling was between 0.12 and 0.18, which indicated that the coupling was reasonably well reflected in the maps.

Only 110 of the 148 papers were bibliographically coupled. The majority of the pairs of papers with the largest number of mutually shared references belonged to ML community only, or both to ML and another one. Of LS-only papers that formed part of a coupled pair, all were 'non-pure' LS papers. Due to incomplete references, in several cases author name had to be 'replaced' by a made-up name (the same everywhere), in order for Bibexcel to work properly. Thus several pairs actually 'share' made-up authors. This could be corrected in the future by, for example, using different made-up names.

The MDS program has an upper limit on the number of papers that can be mapped. 62 papers were selected based on the following criteria: all mixed-category papers should be included; there should be an equal amount on papers in each category as far as possible; most frequently coupled papers should be included. Figure 2 shows the result of the mapping. Circle sizes indicate the total number of shared references for each paper, and lines between two papers indicate that they are bibliographically coupled. The papers in the centre have many links with other papers. Those far down have lowest coupling frequencies. The same map is shown in Figure 3, but with the community tags only, and with lined groupings of the three communities. Papers belonging to several communities are left unmarked.

On both maps, ML papers are situated in the upper right corner and towards the centre, with IR papers continuing on their left, whereas LS papers are separated from the two of them and are positioned much lower, because they have lower coupling frequencies. LS papers are also much more scattered throughout the area, and connected with fewer lines to others because their coupling links are rarer. One can see that ML and IR are more

closely related to each other than to the LS community. ML, and then the IR community, are most frequently coupled ones. Those LS papers with more links to IR and ML papers and with higher coupling frequencies belong to the 'non-pure' subgroup. The majority of mixed category papers are positioned close to either of the categories they were assigned to, indicating which group they belong to more.

Most clearly seen in Figure 3, the three communities form more or less distinct groupings. By further examining LS papers positioned between ML and IR areas, it was discovered that those were papers from the subgroup of LS coming from ML or IR but using a manually created vocabulary. This shows that even the group using controlled vocabularies couples with ML and/or IR, and not with other, 'pure' LS papers.

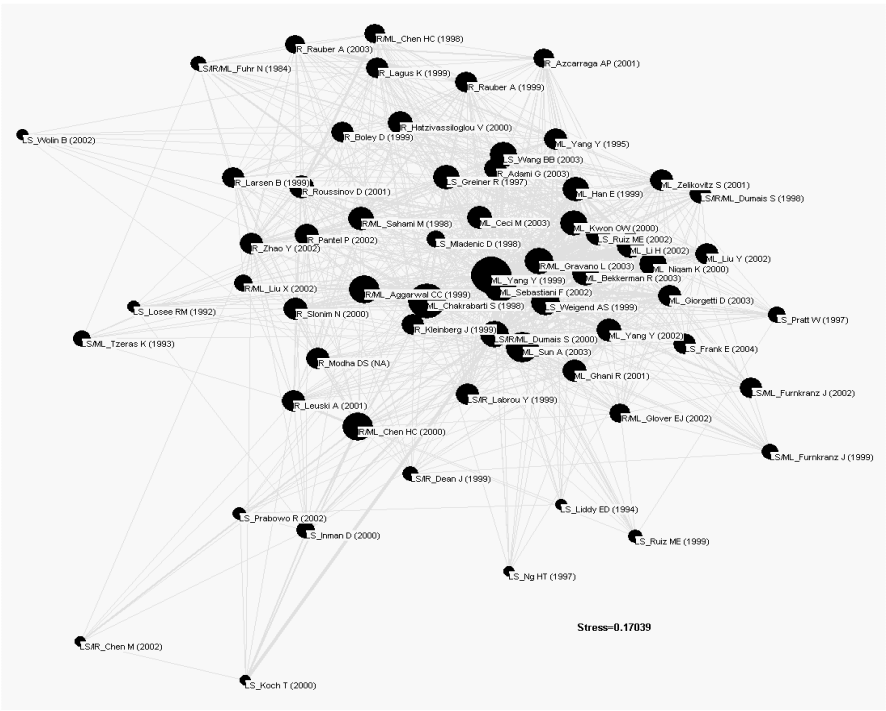


Figure 2: Bibliographic coupling map based on 62 selected papers, with circle sizes indicating the number of shared references

5 Conclusion

Using simple bibliometric methods on the sample of 148 papers, our hypothesis, that the three different communities researching automated classification do not communicate to a large extent, has been confirmed. Absence of ideas exchange was especially the case for the LS community, whereas the ML and IR community exchange ideas to a certain degree. The study of direct links showed that there was not a single case where 'pure' LS authors in the sample were cited by either of the two other communities. The situation was the same the other way around. ML and IR cited each other more but in many cases the authors citing another community's authors, themselves belonged to another community as well. Based on the bibliographic coupling analysis, one can see how the three communities form more or less distinct groupings. One could also see that ML and IR more closely related to each other than to the LS community. The LS and IR community were also most frequently coupled ones. It was discovered that those papers from the subgroup of LS coming from ML or IR but using a manually created vocabulary coupled with ML and/or IR, and not with other, 'pure' LS papers.

Further research would be based on a bigger sample and would deal, e.g., with the following questions: changing trends throughout different periods, and a more detailed analysis of why direct and indirect links are lacking between LS and the other two communities, in spite of appearance of ML and IR papers that employ controlled vocabularies.

Acknowledgements

The Swedish Agency for Innovation Systems and the Danish Ministry of Culture (grant no. A2004-06-028) has in part provided funding for this study. The authors wish to thank prof. dr. sc. Wolfgang Glänzel for his comments on the paper, given within a course organized by Nordic Research School in Library and Information Science (NORSLIS).

Reference

Kessler, M. M. (1963). An experimental study of bibliographic coupling between technical papers. *IEEE transactions on information theory,* 9(1), 49-51.

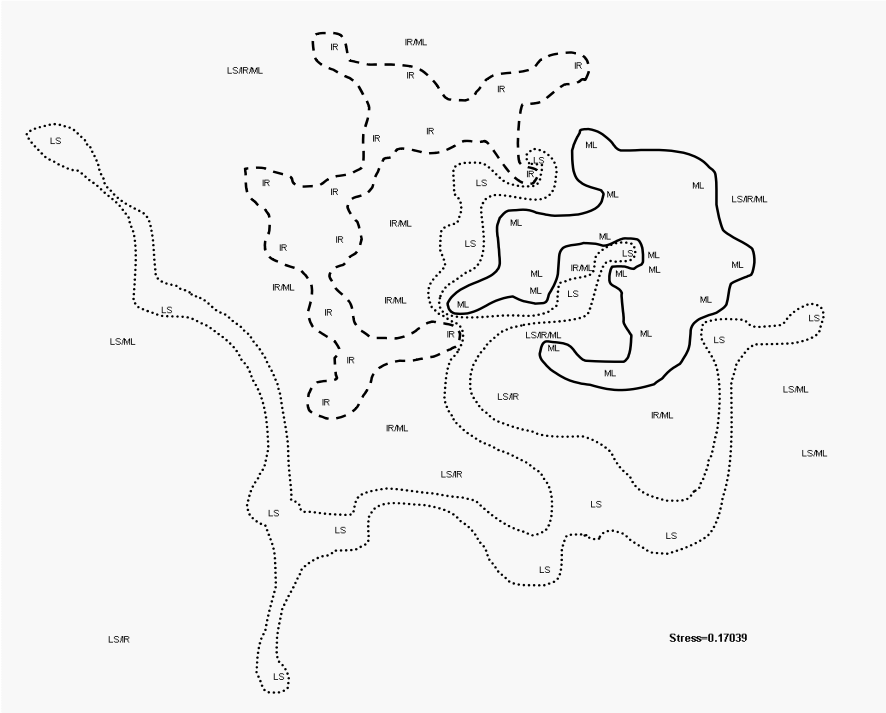


Figure 3: Bibliographic coupling map based on 62 selected papers with groupings emphasised

III

Users Browsing Behaviour in a DDC-Based Web Service: A Log Analysis

Abstract. This study explores the navigation behaviour of all users of a large web service, Renardus, using web log analysis. Renardus provides integrated searching and browsing access to quality-controlled Web resources from major individual subject gateway services. The main navigation feature is subject browsing through the Dewey Decimal Classification (DDC) based on mapping of classes of resources from the distributed gateways to the DDC structure.

Among the more surprising results are the hugely dominant share of browsing activities, the good use of browsing support features like the graphical fish-eye overviews, rather long and varied navigation sequences, as well as extensive hierarchical directory-style of browsing through the large DDC system.

1 Introduction

As many research communities are increasingly concerned with issues of interaction design, one of the current foci in information science is on user behaviour in seeking information on the World Wide Web. A frequently applied methodology for studying this behaviour is log analysis. This approach has several advantages: users do not need to be directly involved in

the study, a picture of user behaviour is captured in non-invasive conditions, and every activity inside the system can be tracked.

User log studies mainly use the average analytical approaches of existing software packages for statistical reporting. Such software provides limited knowledge of user behaviour[1], since it only produces comparatively general insights into aspects of information services, such as number of users per month or the mostly followed hyperlink, and thus tells little about specific navigation behaviour.

A variety of aspects of user information-seeking behaviour using log analysis have been studied previously, in digital libraries[2], web search engines[3,4,5], and other web-based information services. Browsing behaviour has not been studied that much.

The common belief seems to be that users prefer searching to browsing: Lazonder[6] claims "…students strongly prefer searching to browsing". Jacob Nielsen[7] states the following: "Our usability studies show that more than half of all users are search-dominant, about a fifth of the users are link-dominant, and the rest exhibit mixed behaviour. The search-dominant users will usually go straight for the search button when they enter a website: they are not interested in looking around the site; they are task-focused and want to find specific information as fast as possible. In contrast, the link-dominant users prefer to follow the links around a site: even when they want to find specific information, they will initially try to get to it by following promising links from the home page. Only when they get hopelessly lost will link-dominant users admit defeat and use a search command. Mixed-behaviour users switch between search and link-following, depending on what seems most promising to them at any given time but do not have an inherent preference".

This has had implications for building searching-oriented user interfaces. However, those results could be dependent on a number of issues that might have not yet been recognized. One such issue is, for example, the role of the web page layout in "favouring" either of the two strategies. Hong[8] conducted a study on browsing strategies and implications for design of web search engines. The study reports that existing browsing features of search engines are insufficient to users. Even within the Renardus project, an initial belief about potential user requirements was that end-users preferred searching to browsing[9]. After the browsing interface has been built, it showed that browsing was much favoured.

The overall purpose of our project was to gain insights into real users navigation and especially browsing behaviour in a large service on the web. They could be used to improve such services, in our case the Renardus

service[10] which offers a large DDC browsing structure. It is a distributed web-based service which provides integrated searching and browsing access to quality controlled web resources from major individual subject gateway services across Europe (Renardus was funded by the EU's Information Society Technologies 5th Framework Programme until 2002).

The research aimed at studying: the unsupervised usage behaviour of all Renardus users, complementing the initial Renardus user enquiry; detailed usage patterns (quantitative/qualitative, paths through the system); the balance between browsing, searching and mixed activities; typical sequences of user activities and transition probabilities in a session, especially in traversing the hierarchical DDC browsing structure; the degree of usage of the browsing support features; and typical entry points, referring sites, points of failure and exit points. Because of the high cost of full usability lab studies, we also wanted to explore whether a thorough log analysis could provide valuable insights and working hypotheses as the basis for good usage and usability studies at a reasonable cost.

The remainder of the paper provides a short background information about Renardus (I. Background); the methodology applied in this study is described in section two (II. Methodology); the analysis, hypotheses and results regarding the general usage of Renardus, the browsing behaviour and the usage of the DDC are presented in the third section (III. Results). A summary of the results and some ideas for further investigation conclude the paper (IV. Conclusion).

2 Background

2.1 Renardus service

Renardus[10] exploits the success of subject gateways, where subject experts select quality resources for their users, usually within the academic and research communities. This approach has been shown to provide a high quality and valued service, but encounters problems with the ever increasing number of resources available on the Internet. Renardus is based on a distributed model where major subject gateway services across Europe can be searched and browsed together through a single interface provided by the Renardus broker. The Renardus partner gateways cover over 80 000 predominantly digital web-based resources from within most areas of academic interest, mainly written in English.

The Renardus service allows searching several Subject Gateways simultaneously. What is searched are "catalogue records" (metadata) of quality controlled web resources, not the actual resources. There are two ways to search the service, either through a simple search box that is available on the Renardus "Home" page or the "Advanced search" page allowing combination of terms and search fields and providing options to limit searches in a number of different ways. A pop-up window of a list of words alphabetically close to the entered word (for title, DDC, subject and document type) supports the search term selection.

Apart from searching, Renardus offers subject browsing in a hierarchical directory-style (cf. e.g. [14]). It is based on intellectual mapping of classification systems used by the distributed gateway services to the DDC. There are also several browsing-support features. The graphical fish-eye display presents the classification hierarchy as an overview of all available categories that surround the category one started from, normally one level above and two levels below in the hierarchy. This allows to speed up the browsing and get an immediate overview of the relevant Renardus browsing pages for this subject. The feature "Search entry into the browsing pages" offers a short-cut to categories in the browsing tree where the search term is occurring. The lower half of the browsing pages, as a result of the classification mapping, offers the links to the "Related Collections" of the chosen subject. In case users do not want to jump to the parts of the gateways offering related collections, an option of Merging the resource-descriptions from all related collections is available.

For a more detailed description of Renardus, see, for example Koch, Neuroth, and Day[11]. All related publications are given at the web page "Project Archive and Associated Research and Development"[12].

3 Methodology

Before Renardus was finally released and the EU project concluded in 2002, an end user evaluation of the Renardus pilot subject gateway[13] was carried out during Fall 2001 which led to some service improvements. The results and shortcomings of this initial user study stimulated us to try the full study of Renardus user logs which is presented in this paper.

Log analysis was chosen because it costs considerably less than full usability lab studies and has the advantage that it is an unobtrusive means of capturing unsupervised usage. This thorough log analysis required several steps which are described below: cleaning of the log files, defining of user

sessions, categorization into activity types and the creation of datasets and structures to allow the creation of statistics and the testing of hypotheses.

3.1 Cleaning the log files

The log files used spanned 16 months between Summer 2002 and late Fall 2003.

They first had to be cleaned from entries created by search engine robots, crackers, local administration, images etc. The largest group of removed entries, almost half of all log entries, was that containing images and style sheets (1 107 378). Further, 516 269 entries were removed because they originated from more than 650 identified robots, and an additional 12 647 entries because they were from crackers. Various other entries not relating to real usage of Renardus for information seeking, e.g. 17 586 redirections, about 9 000 local administrative activities, error codes and HTTP head entries, had to be removed.

Thus, in the first step, the total number of 2 299 642 log entries was reduced to 631 711 entries.

From this dataset only some general Renardus usage statistics was derived. For the analysis of real user behaviour in Renardus several further steps and separate datasets were required.

3.2 Defining sessions

After cleaning the log all entries were grouped into user sessions. A session was heuristically defined as containing all entries coming from the same IP address and a time gap of less than one hour to the prior entry from the same IP-number.

3.3 Defining activity types

Each log entry was classified into one of eleven different main activities offered by Renardus. These activities were then used to characterize user behaviour, via a typology of usages and sequences of activities.

Browsing activities:

'Gen. Browse', hierarchical directory-style browsing of the DDC (cf. e.g. [14]); 'Graph. Browse', graphical fisheye presentation of the classification hierarchy (cf. e.g. [15]); 'Text Browse', text version of the graphical fisheye presentation; 'Search Browse', search entry into the browsing structure; 'Merge Browse',

merging of results from individual subject gateways; 'Browse', DDC top level browsing page on the home page.

Searching activities:

'Simple Search' with 'showsimpsearch' for result display; 'Adv. Search', advanced search with 'showadvsearch' for result display and 'scan' for scanning certain data indices.

Other activities:

'Home Page'; 'Help'; 'Other' other informational pages, including project documentation.

3.4 Creating datasets for studying information-seeking behaviour

In order to try to make sure that we only study human behaviour in Renardus, we removed, in a further step, another 82 490 entries judged as probable machine activities, based on a couple of heuristic criteria, for example, all sessions containing only one entry as well as sessions shorter than two seconds.

Most of the analysis in this paper talking about human activities in Renardus is, thus, based on a dataset containing 464 757 entries grouped into 73 434 user sessions. Only in a few calculations in this paper (especially in the section "Browsing sessions") we use a further subset of this dataset.

The different datasets were stored in a relational database and SQL has been used to query them to create statistical tables and to test various hypotheses against the log file data.

4 Results

4.1 Global usage

Renardus was accessed from 99 605 unique machines (IP-numbers) during the 16 months period studied. With 351 unique top-level domains or countries identified (a considerable part of the IP-numbers could not be identified), Renardus has a truly global audience. IP-numbers from the USA topped the list with about 30%, other .net and .com domains followed with 8-10%, Project partner countries were led by Finland with 5%. Canada, Australia, the Philippines, Italy and India were other countries exceeding 1% of the IP-numbers.

The user sessions are of considerable length: 33% are longer than 2 minutes and still 10% are longer than 10 minutes. The time users might have been exploring participating gateways after leaving Renardus is not included.

The figures indicate that more than 851 different hosts referred users to Renardus. As much as 56% of all referred sessions came from various Google servers and 24% from Yahoo!

Renardus seemed to be able to attract and keep many "faithful" users already during the first 16 months after release. 13% of all unique user machines were returning to the service, which is a comparatively good value.

4.2 Information seeking activities

4.2.1 Main activities, transitions

Figure 1 illustrates the share of each activity and transition in the following ways: the share of each of the main activities is indicated by the circle size; and the share of the major transitions between different activities is indicated by arrow size. Only values above 1% are displayed.

It shows that 60% of all Renardus activities are directory-style browsing using the DDC structure (Gen. Browse; for the abbreviations here and in the following cf. the description under II Methodology: Defining activity types). 48% of all transitions in Renardus are steps from one such topical page/DDC class to another.

The four special browsing support features are comparatively well used. As many as 45% of the sessions dominated by browsing use two or more different types of browsing activities. As many as 14% use three to five different types (see Browsing sessions below).

Use of the graphical DDC browsing overview (Graph. Browse) is the second most frequent activity in Renardus (7%), after the directory-style browsing. The transition from the dominant directory browsing in the DDC structure to a graphical display is clearly the largest single transition in Renardus, after subsequent directory browsing steps.

Related to Gen. Browse, in 11% of the cases, directory-style browsing has been followed by the usage of the graphical overview (see Figure 2).

For further reasoning about these findings see below.

Figure 2 illustrates another important finding: Users tend to stay in the same feature and group of activities, whether it was a single activity like Gen. Browse or a group like browsing, searching or looking for background information, despite the provision of a full navigation bar on each page of the

Renardus service. In particular, the transitions between browsing and searching activities are less frequent than expected and hoped for. Figure 2 demonstrates this by displaying the main transitions from each feature to other features of the service (the percentages displayed close to the arrows relate to the feature they originate from): e.g. 77% of all transitions from one Gen. Browse activity are directed to another Gen. Browse activity and 11% to Graph. Browse.

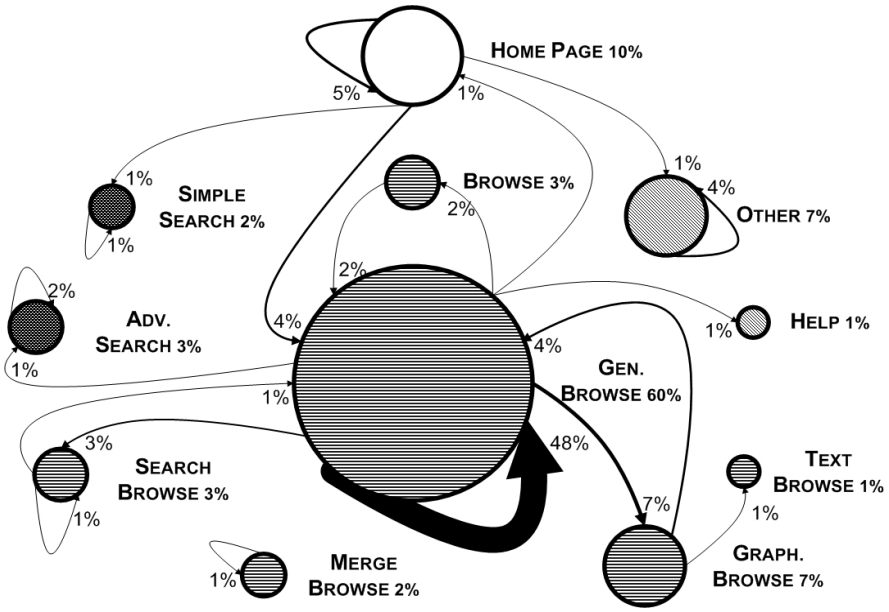


Figure 1: Main Renardus features, indicating their share in all activities, and major transitions between the activities.

As the early user study 2001 showed (in its table 18)[13], the Renardus pilot service was mostly considered very easy or easy to navigate already, although a fifth of the respondents found navigating through the different parts of the service difficult or very difficult.

A conclusion is that advanced online services need to provide some kind of search strategy support. They need to be designed for receiving the user where he/she first enters the system and to assist users navigation through the whole system with more than a ubiquitous navigation bar (which is offered by Renardus on all pages).

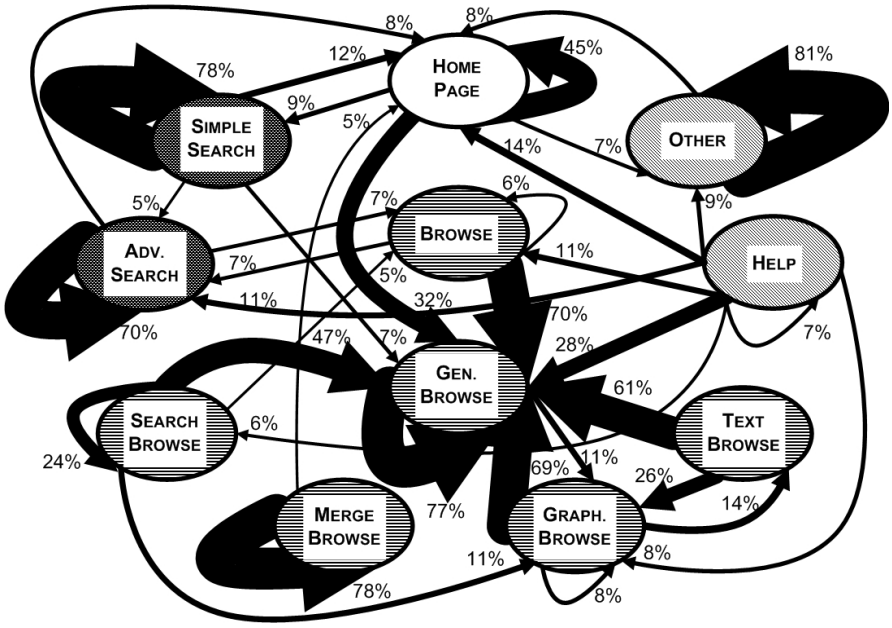


Figure 2: Transition probabilities (more than 5%-transitions only).

4.2.2 General navigation sequences

Many users engage in several different activities during their session: about 46% in one, 20% in two, 16% in three different activities. About 18% of the user sessions have between 4 and 11 different activities.

Many users employ a surprisingly rich variety of navigation and browsing sequences and often alternate between many different features. For example, one session has the following sequence (the first number indicates the number of immediate repetitions of the same feature, the second gives the length of this activity in seconds):

home 3 3; genbrowse 4 31; graphbrowse 1 1; genbrowse 3 17; home 1 1 ; browse 1 1; genbrowse 2 3; searchbrowse 1 1; genbrowse 7 152; searchbrowse 1 1; genbrowse 4 24; help 1 1; home 1 1; genbrowse 4 29; graphbrowse 1 1; searchbrowse 1 1; genbrowse 1 1; searchbrowse 1 1; genbrowse 1 1; browse 1 1; genbrowse 3 2; graphbrowse 1 1; genbrowse 2 2; home 2 2

When we look at the most frequent sequences of activity types (immediate repetition of the same type not counted), we find 4 810 different such sequences with the following top ten (Table 1):

Type of activity	**Sessions**	**%**
(repetitions of) genbrowse	30606	41,7%
home, html, and other	7403	10,1%
genbrowse-graphbrowse-	3860	5,3%
genbrowse-graphbrowse	3590	4,9%
genbrowse-searchbrowse	2812	3,3%
(repetitions of) mergebrowse	2391	3,3%
(repetitions of) showsimpsearch	1705	2,3%
genbrowse-browse-genbrowse	1635	2,2%
genbrowse-searchbrowse-	1236	1,7%
genbrowse-browse	1035	1,4%
all less frequent sequences	17161	23,8%

Table 1. Most frequent sequences of activity types.

The clearly most frequent sequences, apart from mergebrowse and showsimpsearch, are (in and) between browsing activities.

If we look at a more detailed table of sequences including immediate repetitions of the same activity (not reproduced here), the dominance of browsing and the very high number of variations in navigation is well illustrated:

In 73 434 user sessions we find as many as 16 377 different sequences; the top 10 most frequent sequences (with more than 1000 instances each) cover, however, 41,7% of all sessions. In the top 6 and number 9-11 among the 11 most frequent sequences the user does exclusively repeat the same activity. Only no. 7 and 8 involve a switch between different activities (from genbrowse to graphbrowse and from genbrowse to searchbrowse). In the five most frequent cases genbrowse is the repeated activity.

The sequences where only the same activity type is repeated cover about 50% of all sessions.

This further underlines our earlier finding, that a surprisingly large part of the users stay in the same (group of) activities.

4.2.3 Browsing vs. searching

The levels of usage of the main Renardus features are highly uneven (cf. Figure 1). The most surprising finding is the clear dominance of browsing activities: about 80% (dependent of how exactly "dominance of browsing" is defined: 76% of all activities are browsing, 80,5% of all sessions are dominated by browsing). Searching has a share as low as between 3 and 6%.

This is a highly unusual ratio compared to other published evaluations and common beliefs (cf. Introduction). Among possible reasons are:

a) most of the browsing pages are indexed by search engines; 71% of the users reached browsing pages directly via search engines and start their Renardus navigation at a browsing page. These facts "favour" browsing, taken together with the clear tendency to stay in the same (group of) features.

b) the layout of the home page invites browsing by putting the browsing structure on top of the search box. Still, among users starting at the home page, 57% browse and only 12.5% search (only 22% of all users enter Renardus at the home page/the "front door" of the service, however).

In spite of the dominance of browsing and the tendency to stay in the same group of activities we see a certain amount of switching between browsing and searching during the same session:

In as few as 7,3 % of all sessions users switch between a browse and a search activity, out of which 4,5% of sessions have one switch, 1,9% have two, 0,4 have three, and 0,5% have more than three switches.

The largest number of such switches per session is 20. Out of 27 different kinds of switches between browsing and searching, 7 start with a search. Switching from browsing to searching is much more frequent than the opposite. Users at the search pages need to be pointed to the benefits of browsing.

4.2.4 Browsing sessions

For the calculations in this section we use a subset of our usual dataset, containing 378 267 entries in 58 954 user sessions, defined by a share of more than 50% browsing activities: sessions where "browsing is dominant".

The shares of sessions with a certain number of different activities are almost the same as for all Renardus sessions (cf. the beginning of General

navigation sequences). So, even sessions with dominant browsing show as much variety in activities as most other sessions.

Many browsing sessions use more than one type of browsing activity, including the browsing support features Graph. Browse, Text Browse and Merge Browse. As many as 45% of the sessions dominated by browsing show two or more and 14% three to five different types of browsing.

We find up to 95 individual browse activities per session, with gracefully degrading numbers from two activities and down.

4.2.5 Two different groups of users

Because of the big influence of referrers like search engines, 71% of the human user sessions start at browsing pages pointed to them by referrers, 22% start at the homepage (16 300 out of 73 434 sessions). This, at least quantitatively, surprising result stimulated us to check out if these two "groups" of users show significantly different navigation behaviour. Sessions starting at home have almost twice as many entries per session than sessions starting elsewhere (10 vs. 5,8 entries per session; 35,8% of all entries). Thus, home starters carry out many more activities per session than the other user group.

Type of activity	Starting at home		Starting elsewhere	
	Entries	**%**	**Entries**	**%**
Browsing	94 215	56,6	259 471	87,0
Searching	20 831	12,5	8 099	2,7
Other	51 139	30,9	30 684	10,3
Total	166 503		298 254	

Table 2. Types of activities for the two different groups of users.

Users jumping into the middle of the Renardus service are carrying out browsing activities in 87% of all cases and only 2,7% searching activities (Table 2).

Users starting at the Renardus homepage/"frontdoor" show a level of browsing of almost 57%, and 12,5 % searching. Three times as often they visit Other pages and five times as often search pages compared to the other group. These are probably the users who go deliberately to Renardus, whereas a large part of users starting elsewhere, most often in the browsing

pages, end up there "ignorantly" after a search in a search engine. The latter overwhelmingly stay in the browsing activities.

People starting elsewhere have a much higher percentage of browsing among their activities. Home starters do, however, considerably more browsing activities compared to their share of all sessions: 53,2% of the sessions show more than 11, still 36,8% more than 30 browsing activities.

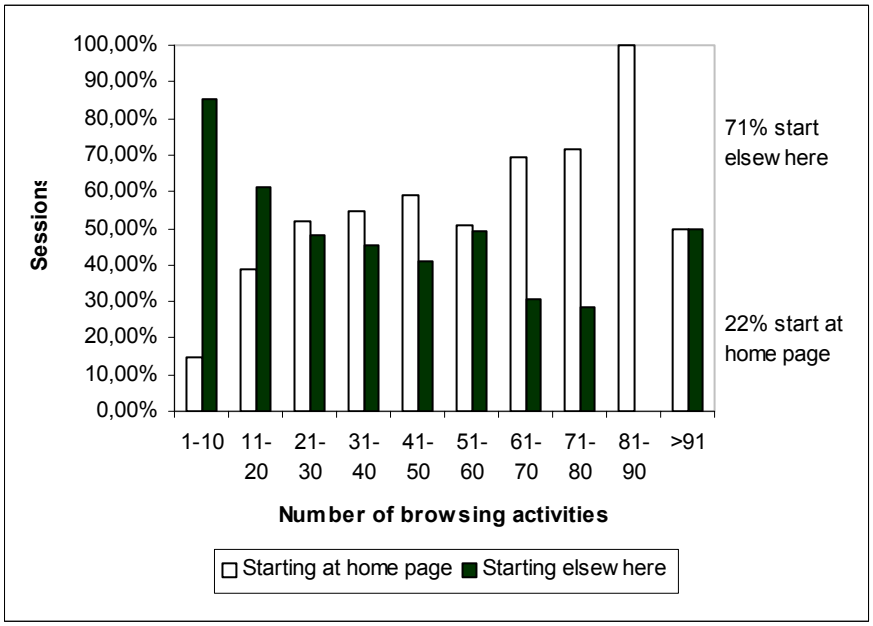


Figure 3: Browsing activities of the two groups of users.

Figure 3 shows that the home starters clearly dominate the sessions with many browsing activities. A more detailed analysis shows that they are active in browsing activities to a higher and increasing degree starting with 8 browsing activities, compared with their share in all sessions (21%). Quite the opposite, users starting elsewhere are overrepresented up to the level of nine browsing activities with an ever-decreasing tendency.

Home starters also exceed their share when it comes to the number of different activity types, all types counted (in browsing sessions). The exception is when there are three different activities, strangely enough. From five different activities and higher, they have more than twice their share and dominate clearly.

When it comes to the number of different browsing types (in browsing sessions), home starters exceed their share when it comes to carrying out between three and five different browsing types considerably.

4.3 DDC usage

4.3.1 DDC analysis

Analysis of the popularity of DDC sections and classes and the navigation behaviour of users in the DDC structure allow good insights into the distribution of topical interests and into the suitability of DDC system and vocabulary. The findings from the log analysis can, however, only help create hypotheses and need to be complemented by investigative sessions with the users.

The most frequently used parts of the DDC hierarchy at the top hierarchical level are given in Table 3.

Entries	DDC	Class
50784	3	Social sciences
46209	5	Science
30955	6	Technology
26015	2	Religion
22081	7	Arts & recreation
17994	8	Literature
16828	9	History & geography
16527	0	Computers, information &
13839	4	Language
13428	1	Philosophy & psychology

Table 3. Most frequently used parts of the DDC hierarchy at the top hierarchical level.

All DDC classes show generally good usage levels (users just jumping to one class and not continuing browsing are not counted). Compared to what one would expect in a global internet setting, Religion ranks surprisingly high and Computers etc. unexpectedly low (see Table 3). Here could the vocabulary used in the DDC captions play a certain role, e.g. many

computing-related terms used in Internet searching are not directly occurring in the captions.

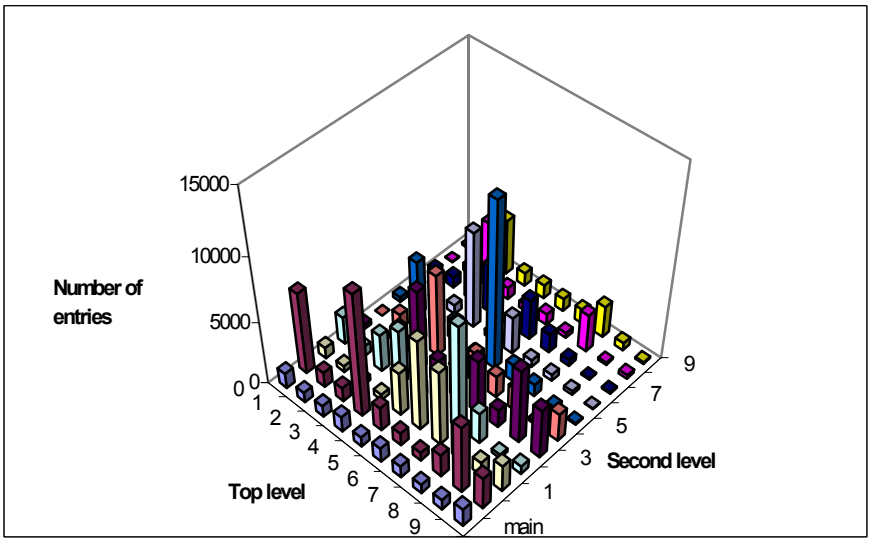


Figure 4: Most frequently used parts of the DDC hierarchy at the second hierarchical level.

On the second hierarchical level, surprisingly large topical areas are Christian denominations (DDC 28), German & related literatures (83), Social problems (36) and Earth Sciences (55; cf. Figure 4).

Unexpectedly frequent visits to individual topics like 552.1 Igneous rocks (the sixth most visited individual page with 2 436 directory browsing activities) can be due to the fact that little information might be found about such a concept in the search engines or to the fact that other sites made prominent links to this topic page in Renardus.

4.3.2 Directory style of browsing in the DDC hierarchy

The directory-style of browsing in the DDC-based browsing structure is the clearly dominating activity in Renardus (about 60%). 67% of all browsing activities are DDC directory browsing (254 660 out of 378 264 entries in browsing sessions). Two thirds of the latter (167 628) appear in unbroken sequences. In these cases, not even browse support features are used between

directory browsing steps. While the clear majority limit themselves to up to 10 such steps (for distribution see the Figure 5), we found surprisingly long unbroken sequences of up to 86 steps in the DDC directory trees.

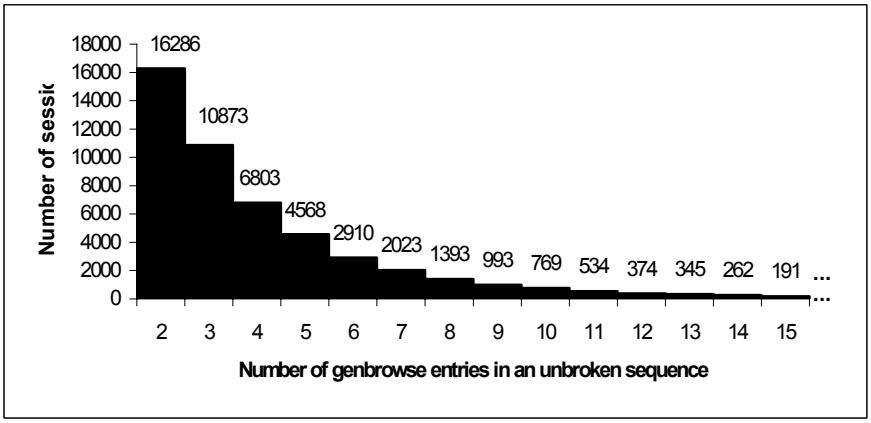


Figure 5. Number of genbrowse activities in sessions (up to 15).

These are very unexpected results. People looking for information on the web are often said doing only very few clicks, switching frequently to other services and activities, having a very short attention span and similar. Browsing the DDC hierarchies in a directory style of steps at such quantity and lengths is one of the most significant results of this log study.

4.3.3 Jumping in the DDC hierarchy

Since the DDC browsing area in the Renardus user interface displays the higher levels in the hierarchy in addition to the "father" and the "child" classes, we can find out to what a degree users are doing such jumps in the DDC hierarchy during unbroken directory browsing sequences.

Two of the support features, the graphical overviews and the "search entry to browsing pages", were designed to relieve users from the "pain" of having to jump around in the hierarchy. Jumping one step up and another step down in the directory-style display is probably faster and easier than using the support features, moving farther away would possibly have been easier using the support features.

Here is an example of a session featuring jumps within unbroken directory browsing:

start 62-; go to 624; go to 624.1; jump to 62-; go to 625; go to 625.1; go to 625;
go to 62-; go to 627; jump to 628; go to 628.1

20,2% of all steps in sessions featuring unbroken directory browsing are jumps.

Jumps occur in 40,8% of these sessions. In the sessions with jumps, on average 1,7 jumps are carried out.

This is a decent number of cases but not excessively high. Many users have used the support features, especially the graphical overviews and, thus, not the jumping in the directory. It indicates at least, that the necessity to jump in the hierarchy is not off putting users.

As seen from Figure 6 the probability for a user in one session to browse in several main DDC classes increases with the length of the session. This might seem natural but it also implies that the longer the session, the shorter time spent within one main DDC class before moving to another. Each point in the figure is based on several sessions that together contain more than 2 000 browsing entries. Due to the heavy dominance of shorter sessions, the overall mean probability of moving between DDC main classes in a session is 3%.

Figure 7 shows a few individual sessions plotted with the number of browsing steps versus the visited DDC classes. For example, all classes within the '1--' branch of DDC are displayed between 100 and 200 on the vertical axis in such a way that the hierarchy is preserved, e.g. the closer two classes are in the hierarchy the closer they are plotted in the figure. Thus a horizontal line indicates that the user stays within a narrow area of DDC while vertical parts indicate jumps between different areas of DDC. A 'G' indicates that the graphical overview was used while an 'S' indicates that the search entry to the browsing structure was used at the indicated points in the sequence.

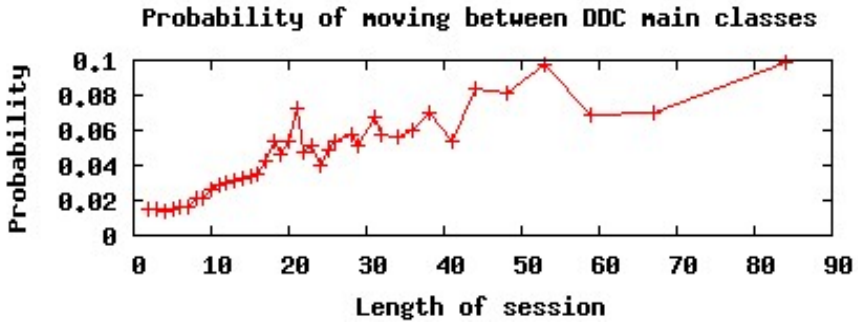


Figure 6: Probability of moving between DDC main classes.

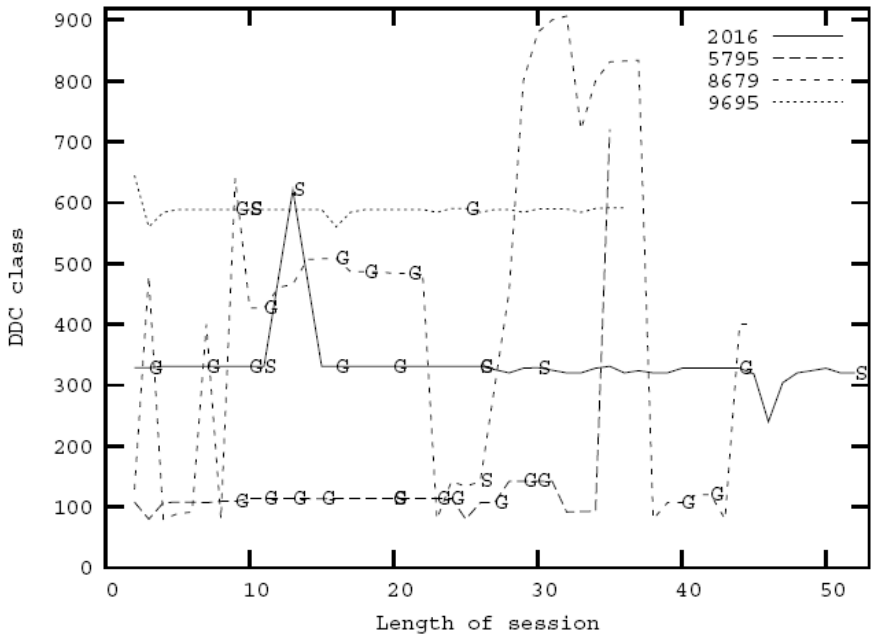


Figure 7. DDC browsing behaviour per session.

4.3.4 Keywords and browsing

In order to get an indication whether the user managed to come close to his/her topic on the DDC browsing pages, we compared keywords put into a search engine respectively into Renardus Search with the browsing pages visited.

Here are some examples of keywords entered into Google and the => Renardus DDC class the user selected from the search result:

ancient continents => History of ancient world; of specific continents, countries, localities; of extraterrestrial worlds
perspective drawing => Drawing & decorative arts
"statistics of south america" => General statistics of specific continents, countries, localities in modern world
writing systems and etymology => Standard language--description and analysis
kinds of sedimentary rocks => Specific kinds of rocks

The sample studied showed very good hits in the Renardus DDC pages. Most queries had good hits in the DDC caption (which is also used as the title of the page), about 13% of the cases had partial hits there and partial in other class and directory "titles" mentioned on the same page (father, child DDC classes; names of mapped directories from cooperating subject gateways). Most successful questions used 2-3 query terms; only 3% used one term.

It seems that a good initial hit is required to invite users to continue browsing in Renardus (the data here is derived from sessions containing more than one activity).

The result says more about the search engines (Google in the case of our sample) ranking algorithm than about the Renardus pages and the suitability of the DDC captions. Predominantly hits with several search terms in the title and top half of a browsing page have a chance to appear in the top of Google search results on most terms.

When checking queries and hits in Other Renardus pages (background and project information), we found great results too: most hits seemed relevant and we couldn't find many wrong hits on topical questions.

Here are examples showing search terms entered into Renardus Search (Q:) and DDC classes/pages used during browsing, in each case gathered from the entire session (sessions starting with Search and continuing with Browse activities only. Queries and DDC captions are separated by ';'):

Q=chopin; vieuxtemps;
DDC=Arts & recreation; Music; Composers and traditions of music;

Q=paperin+valmistus; paperin+valmistus; papermaking; paper+technology;
DDC=Technology; Engineering; Engineering of railroads and roads; Engineering of railroads and roads; Engineering of railroads and roads; Railroads; Railroads; Astronautical engineering; Technology; Engineering; Engineering and allied operations; Engineering mechanics and materials; Science; Chemistry; Chemistry; Organic chemistry; Technology; Chemical engineering; Chemical engineering and related technologies; Biotechnology; Biotechnology; Pulp and paper technology; Genetic engineering; Electrical engineering; lighting; superconductivity; magnetic engineering; applied optics; paraphotic technology; electronics; communications engineering; computers; Electrical engineering; lighting; superconductivity; magnetic engineering; applied optics; paraphotic technology; electronics; communications engineering; computers; Electronics; Special topics; Optoelectronics; Pulp and paper technology; Conversion of pulp into paper, and specific types of paper and paper products; General topics; Properties, tests, quality controls;

The results of the evaluations of our sample remind us that users very well do follow more than one topic of interest during one session in an information system. In our sample 70% of all sessions seemed to pursue one topic, 23% two topics, 2% three topics and 5% seemed to browse around without specific question. In some cases, topics looked for in Renardus Search are not pursued when browsing, in other cases, a new topic (most often one) is investigated after the switch to browsing.

5 Conclusion

The main purpose of this study was to explore the navigation behaviour of all users of a large web service, Renardus, using web log analysis, in order to improve the user interface and, especially, the browsing features of the system. In addition, we aimed at gaining some more general insights into users browsing and navigation in large subject classification structures, the benefits from system support and the problems and failures occurring.

Our study indicates that a thorough log analysis can indeed provide a deeper understanding of user behaviour and service performance. Being an unobtrusive means of capturing unsupervised usage and offering a complete and detailed picture of user activities it can reveal quantitatively comprehensive, sometimes unexpected results, far beyond plain statistics.

In contrast to common belief, our study clearly indicates that browsing as an information-seeking activity is highly used, given proper conditions. About 80% of all activities in Renardus are browsing activities. A contributing reason to that dominance is the fact that a very high percentage (71%) of the users are referred from search engines or other linking sites directly to a browsing page in Renardus. The layout of the home page "invites" browsing, which certainly contributes to the fact that even users starting at the home page predominantly use the browsing part of the service.

Our study leads to a hypothesis which deserves further research: browsing is perceived as useful and dominates navigation in services similar to Renardus and under proper conditions.

The good use of the browsing support features, especially graphical overview and search entry to browsing pages, suggests that it would be worthwhile to further develop such support.

Since most visitors jump into the middle of the service, there might be a need to redesign the browsing pages so they would better serve as full-fledged starting points for comprehensive Renardus exploration. The ubiquitous navigation bar seems not sufficiently inviting. In making such

changes, it would also be important to better understand the details of site indexing and ranking algorithms in search engines.

The study of navigation sequences shows that users employ a rich variety of navigation and browsing sequences, including rather long and highly elaborated paths through the system. Nevertheless, quantitatively dominating is, to a quite surprising degree, the tendency to stay in the same group of activities or individual activity, whether browsing, searching or background information. This points us to the importance of providing "search strategy" support to the users at the page where their actions take place.

From the behaviour as documented in the log files we could identify two clearly different groups of users: people starting at the homepage/frontdoor of the service (22%), and the majority of the users starting elsewhere. There are dramatic differences in their activity in the service. People starting at the homepage show almost twice as many activities per session, and use the non-browsing features three to five times as often. Their share of the browsing activities is smaller, but they primarily engage in the long sequences of browsing activities (8 and longer) and employ more different types of browsing and more different types of other activities in a session. The home page starters are seemingly a minority but represent high quality of usage of the service in a way the system designers have imagined and intended.

The DDC directory browsing is the single clearly dominating activity in Renardus (60%). Two thirds of it is done in unbroken directory browsing sequences. We see a surprising average and total length of such browsing sequences, opposing the common belief of the short attention span of users online.

Thus, we get the surprising hypothesis that sequential, directory style of hierarchical (classification) browsing is found popular and useful in large services like Renardus, especially when there is graphical support.

Comparisons between search terms used and topics browsed indicated a very good chance to get relevant results from Renardus browsing when more than one search term was used. People using Renardus Search were capable to find browsing pages corresponding to their queries. The system invited certainly to pursue more than one topic during a session.

6 Future work

Our findings indicate that log analysis has a clear potential as a method for studying information behaviour and the proper design of information systems. A lot could be gained from future work to investigate questions such as:

- To what degree does the actual design of the system influence user behaviour, especially with regard to the difference in usage levels of browsing versus searching activities?
- Can we identify further specific usage and browsing patterns and different behaviours of specific user groups?
- What is the influence of the use of end-user adapted and multilingual DDC captions on users browsing behaviour?
- How can we provide search strategy support and further improve the support for systematic browsing of large subject structures?
- What is the importance of the details of site indexing in search engines for the discovery of and navigation in large browsing systems?
- How can pages be redesigned so that they better serve as full-fledged starting points?

For more important results and improvements one would need to go beyond the log analysis and:

- evaluate user behaviour in supervised sessions/usability lab
- evaluate the accuracy and success of Renardus to help answering user questions
- use local URLs to identify what pages outside Renardus users explore as a result of Renardus navigation (links to participating subject gateways).

7 Acknowledgements

The Swedish Agency for Innovation Systems provided the main funding for this research. This work was partially funded by EU under project ALVIS – Superpeer Semantic Search Engine (EU 6. FP, IST-1-002068-STP). This work was partially funded by DELOS – Network of Excellence on Digital Libraries (EU 6. FP IST, G038-507618).

References

1. Harry Hochheiser and Ben Shneiderman. "Using interactive visualizations of WWW log data to characterize access patterns and inform site design." *Journal of the American Society for Information Science and Technology* 52, no. 4 (2001): 331–343.

2. S. Jones, S. J. Cunningham, R. McNab, and S. Boddie. "A transaction log analysis of a digital library." *International Journal on Digital Libraries* 3, no. 2 (2000): 152-169.

3. C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. "Analysis of a very large web search engine query log." In: *SIGIR Forum,* 33, no. 1(1999): 6-12. http://doi.acm.org/10.1145/331403.331405

4. Seda Ozmutlu, Amanda Spink, and Huseyin C. Ozmutlu. 2004. "A day in the life of web searching: an exploratory study." *Journal of Information Processing and Management* 40, no. 2 (2004): 319-345. http://dx.doi.org/10.1016/S0306-4573(03)00044-X

5. S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. "Hourly analysis of a very large topically categorized web query log." In: *Proceedings of the 27th annual international conference on Research and development in information retrieval, Sheffield, United Kingdom*, 2004, 321-328. http://doi.acm.org/10.1145/1008992.1009048

6. Ard W. Lazonder. "Principles for Designing Web Searching Instruction." *Education and Information Technologies* 8 (June 2003): 179–193. P. 181.

7. Jacob Nielsen. "Search and You May Find." Jakob Nielsen's Alertbox for July 15, 1997. http://www.useit.com/alertbox/9707b.html.

8. Xie Hong. "Web browsing: current and desired capabilities." In: *20th Annual National Online Meeting, 18-20 May 1999 , New York, NY, US,* 523-37.

9. User requirements for the broker system: Renardus Project Deliverable D1.2. 2000. P. 23. http://www.renardus.org/about_us/deliverables/d1_2/D1_2_final.pdf.

10. Renardus Home Page. http://www.renardus.org/.

11. Traugott Koch, Heike Neuroth and Michael Day. "Renardus: Cross-browsing European subject gateways via a common classification system (DDC). In: "Subject Retrieval in a Networked Environment", *Proceedings of the IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing and the IFLA Section on Information Technology, 14-16 August 2001, Dublin, OH, USA*, 25-33. München: UBCIM Publications New Series Vol. 25, 2003. Manuscript at: http://www.lub.lu.se/~traugott/drafts/preifla-final.html

12. Renardus Project Archive and Associated Research and Development. 2002. http://www.renardus.org/about_us/project_archive.html.

13. User evaluation report: Renardus Project Deliverable D5.2. 2002. http://www.renardus.org/about_us/deliverables/d5_2/D5_2_final.pdf.

14. Technology:Agriculture : page. http://www.renardus.org/cgi-bin/genDDCbrowseSQL.pl?ID=10191&node=AAZNG

15. Graphical browsing page for Technology ...: Mining for specific materials. http://www.renardus.org/cgi-bin/imageDDCbrowseSQL.pl?node=ABDPH&ID=10193 &pmat=N&pnavnode=Y&pgraph=matcirc

All electronic resources have been accessed 20 January 2005.

IV

Browsing and Searching Behavior in the Renardus Web Service. A Study Based on Log Analysis.[1]

1 Introduction

Renardus (http://www.renardus.org) is a distributed Web-based service, which provides integrated searching and browsing access to quality-controlled Web resources from major individual subject gateway services across Europe (was funded by the EU's Information Society Technologies 5th Framework Program). Navigation features are, among others, simple and advanced search, and subject browsing. Browsing is based on intellectual mapping of classification systems used by the distributed gateway services to the Dewey Decimal Classification (DDC). In addition to the dominating hierarchical directory-style of browsing (Gen. Browse), there are several other supporting features: graphical fisheye presentation of the classification hierarchy (Graph. Browse), search entry into the browsing structure (Search Browse) and merging of results from individual subject gateways (Merge Browse). With the overall purpose of improving Renardus, the research aims to study: the detailed usage patterns (quantitative/qualitative, paths through the system); the balance between browsing and searching or mixed activities; typical

[1] Poster.

sequences of usage steps and transition probabilities in a session; typical entry points, referring sites, points of failure and exit points; and, the usage degree of the browsing support features.

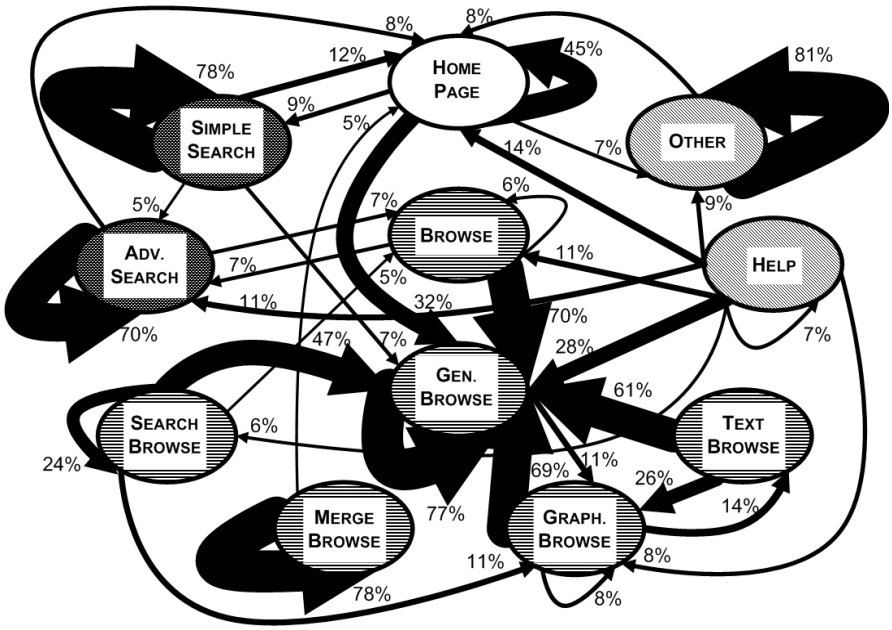
2 Approach

The Renardus project did a limited human evaluation of the service. Because of the high cost of full usability lab studies, we wanted to explore to what a degree a thorough log analysis, catching unsupervised usage, could provide valuable insights and working hypotheses as the basis for good usage and usability studies. Many sources of problems may be discovered at this stage. A thorough log analysis requires several steps, starting with cleaning the log files with regard to activities from search engines, crackers, local administration, images etc. More than 2.3 million Renardus log entries boiled down to 630,000 user entries. The second step, based on heuristics, was to remove further 80,000 entries as probable machine activities. In order to study behavior we grouped log entries into user sessions. The basis for our further analysis were 155,000 user sessions, corresponding to 550,000 log entries, spanning over more than one year. Each entry was classified into one of eleven different activities offered by Renardus. These activities were then used to characterize user behavior, via a typology of usages and sequences.

3 Preliminary findings

The most surprising finding is the clear dominance of browsing activities (80%). Among possible reasons are: a) the fact that 71% of the users reach browsing pages directly via search engines; b) the layout of the home page focuses on browsing. Users tend to stay in the same group of activities, whether it is browsing, searching or looking for background information, despite the provision of a full navigation bar on each page of the service. The following illustration demonstrates this by displaying the main transitions from each feature to other features of the service.

Services like Renardus need to be designed for receiving the user where she first enters the system and provide search strategy support for the full usage of the system's features. The special browsing support features of the service are quite well used and worthwhile to further develop. Many users employ a surprisingly rich variety of navigation and browsing sequences and often alternate between many different features.



Directory-style browsing in the DDC-based browsing structure is the clearly dominating activity in Renardus (60%). We found surprisingly long unbroken sequences of up to 90 steps in the DDC directory trees, even if the vast majority limits themselves up to 10 such steps. The usage of the graphical DDC browsing overview is the second most frequent activity in Renardus, after the directory-style of browsing. In 11% of the cases, directory-style browsing has been followed by the usage of the graphical overview. Systematic browsing of large information systems with the help of classification hierarchies seems to be widely accepted by users, especially when there is graphical support.

These findings indicate that a thorough log analysis can provide deeper understanding of how the service really works and can be improved. They might offer useful hypotheses for advanced user studies. Future work aims at investigating questions like: what influences the different usage levels of browsing versus searching activities?; to what a degree is the actual design of the system influencing user behavior?; which important changes in design are called upon by the results of such user and log studies?; and, how can we

provide search strategy support and improve the support for systematic browsing of large subject structures?

Acknowledgements

Swedish Agency for Innovation Systems provided funding for this research.

V

Log Analysis of User Behaviour in the Renardus Web Service[1]

1 Introduction

Renardus (http://www.renardus.org) is a distributed Web-based service, which provides integrated searching and browsing access to quality controlled Web resources from major individual subject gateway services across Europe (funded by the EU's Information Society Technologies 5th Framework Programme until 2002). Navigation features are, among others, simple and advanced search, and subject browsing. Browsing is based on intellectual mapping of classification systems used by the distributed gateway services to the Dewey Decimal Classification (DDC). In addition to the dominating hierarchical directory-style of browsing (Gen. Browse), there are several other supporting features: graphical fisheye presentation of the classification hierarchy (Graph. Browse), search entry into the browsing structure (Search Browse) and merging of results from individual subject gateways (Merge Browse). Fig. 1 shows the main features, indicating their share of the activities (circle sizes) and transitions (arrow sizes) (only values above 1% are displayed):

[1] Poster.

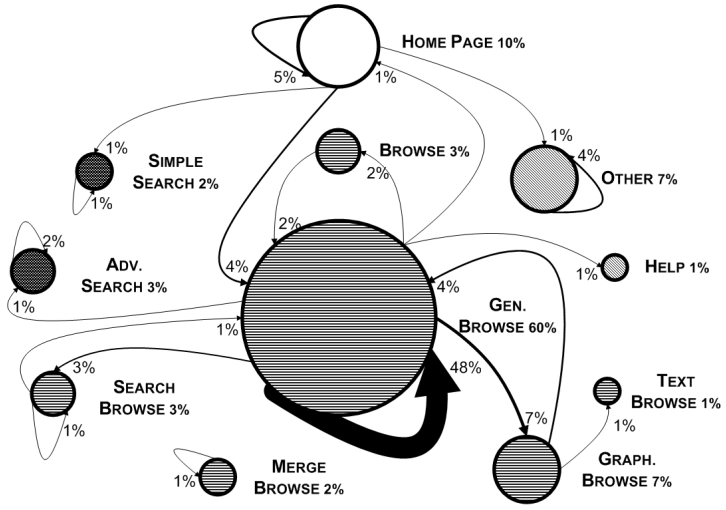


Fig. 1

With the overall purpose of improving the Renardus Web service, the research aims to study:

- the detailed usage patterns (quantitative/qualitative, paths through the system)
- the balance between browsing and searching or mixed activities
- typical sequences of usage steps and transition probabilities in a session
- typical entry points, referring sites, points of failure and exit points
- the usage degree of the browsing support features.

2 Approach

The Renardus project did a limited human evaluation of the service. Because of the high cost of full usability lab studies, we wanted to explore to what a degree a thorough log analysis – monitoring unsupervised usage – could provide valuable insights and working hypotheses as the basis for good usage and usability studies. Many sources of problems might be discovered already at this stage. A thorough log analysis requires several steps, starting with cleaning the log files with regard to activities from search engines, crackers, local administration, images etc. More than 2.3 million Renardus log entries

boiled down to 630,000 user entries. The second step, based on heuristics, was to remove further 80,000 entries as probable machine activities. In order to study behaviour we needed to group log entries into user sessions. The basis for our further analysis turned out to be 155,000 user sessions, corresponding to 550,000 log entries, spanning over the period of 16 months. Each entry was classified into one of eleven different activities offered by Renardus. These activities were then used to characterize user behaviour, via a typology of usages and sequences.

3 Preliminary findings

The log files analyzed show global usage of Renardus from about 99,605 unique machines and 351 unique top-level domains. First figures indicate that about 13% of our unique user machines have been returning to the service, which is a comparably good value for "faithful" users.

The levels of usage of the main Renardus features are highly uneven (cf. Fig. 1). The most surprising finding is the clear dominance of browsing activities (80%). This is a highly unusual ratio compared to other published evaluations and common beliefs. Among possible reasons are: a) the fact that 71% of the users reach browsing pages directly via search engines (Google and Yahoo! dominating); b) the layout of the home page focuses on browsing (22% of all users enter Renardus at the home page/the "front door" of the service).

Users tend to stay in the same feature (e.g. Adv. Search) and group of activities, whether it is browsing, searching or looking for background information, despite the provision of a full navigation bar on each page of the service. Especially the transitions between browsing and searching activities are less frequent than expected and hoped for. Fig. 2 demonstrates this by displaying the main transitions from each feature to other features of the service (the percentages – above 5% – displayed with the arrows relate to the feature they originate from).

Services like Renardus need to be designed for receiving the user where she first enters the system and provide search strategy support for the full usage of the system's features. The special browsing support features of the service are quite well used and worthwhile to further develop. Many users employ a surprisingly rich variety of navigation and browsing sequences and often alternate between many different features. For example, one session has the following sequence (the numbers indicate the repeated usage of the same feature):

home 2 - genbrowse 3 - browse 1 - home 2 - html 3 - genbrowse 6 - graphbrowse 1 - genbrowse 1 - graphbrowse 1 - genbrowse 1 – graphbrowse 1 - textbrowse 1 - graphbrowse 1 - genbrowse 4 - graphbrowse 1 - searchbrowse 2 - graphbrowse 1 - advsearch 1 - graphbrowse 1 - browse 1 - genbrowse 2 - graphbrowse 1 - textbrowse 1 - genbrowse 3 - advsearch 1 - showadvsearch 2 - scan 1 - showadvsearch 1 - scan 2 - advsearch 1 - showadvsearch 1 - browse 1 - genbrowse 1.

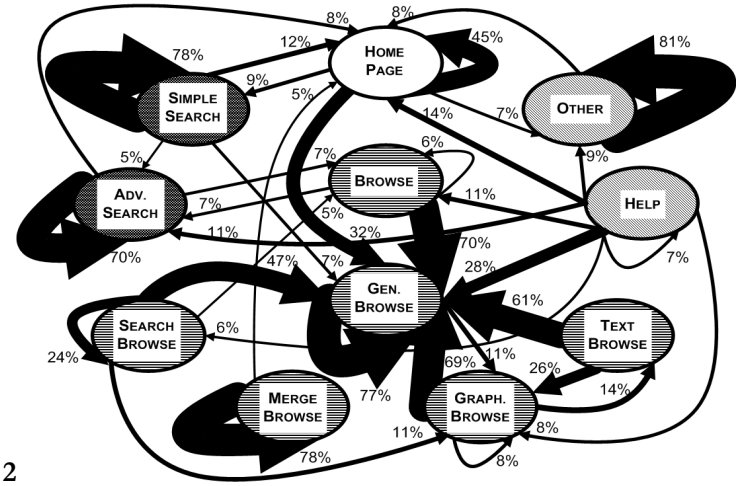


Fig. 2

Directory-style of browsing in the DDC-based browsing structure is the clearly dominating activity in Renardus (60%). We found surprisingly long unbroken sequences of up to 90 steps in the DDC directory trees, even if the clear majority limits themselves up to 10 such steps (cf. the detailed view in Fig. 3b).

Use of the graphical DDC browsing overview is the second most frequent activity in Renardus, after the directory-style browsing. In 11% of the cases, directory-style browsing has been followed by the usage of the graphical overview.

Analysis of the popularity of DDC sections and classes and the navigation behaviour of users in the DDC structure allow good insights into distribution of topical interests and into the suitability of DDC system and vocabulary.

Systematic browsing of large information systems with the help of classification hierarchies seems to be widely accepted by users, especially when there is graphical support.

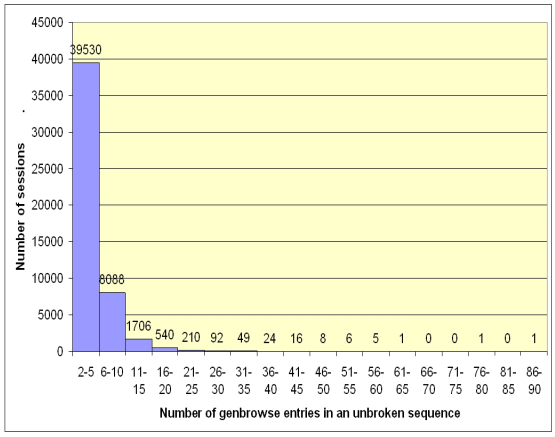


Fig. 3a

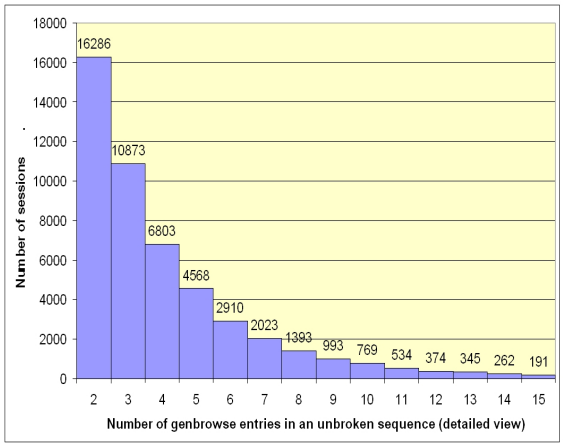


Fig. 3b

4 Future work

These findings indicate that a thorough log analysis can provide deeper understanding of how the service really works and can be improved and they might offer useful hypotheses for advanced user studies.

Future work aims at investigating questions like:

- are there stable usage and browsing patterns and different behaviours of specific user groups?

- to what a degree is the actual design of the system influencing user behaviour, especially with regard to the different usage level of browsing versus searching activities?
- how can we provide search strategy support and improve the support for systematic browsing of large subject structures?

In order to make up for shortcomings of the log analysis approach, the following investigations will be needed:

- use cookies to identify the pages outside Renardus users explore as a result of Renardus navigation
- evaluate user behaviour in supervised sessions/usability lab
- evaluate the accuracy and success of Renardus to help answering user questions.

Acknowledgements

The Swedish Agency for Innovation Systems provided funding for this research.

References

1. Renardus Home Page. http://www.renardus.org.

2. Presentation of some more detailed Renardus log analysis results. http://www.it.lth.se/knowlib/renardus-log/log-analysis.html.

3. Koch, T., Neuroth, H., & Day, M. (2001). Renardus: Cross-browsing European subject gateways via a common classification system (DDC). In *Proceedings of the IFLA Satellite Meeting on Subject Retrieval in a Networked Environment*, 14-16 August 2001, Dublin, OH, USA. UBCIM Publications - New Series Vol. 25, München, 2003, 25-33. Manuscript at: http://www.lub.lu.se/~traugott/drafts/preifla-final.html.

4. Hollman, J., Ardö, A. & Stenström, P. (2002). Empirical observations regarding predictability in user access behaviour in a distributed digital library system. In *Proceedings of the 16th International Parallel and Distributed Processing Symposium*, IEEE, April 2002, 221-228.

VI

Importance of HTML Structural Elements and Metadata in Automated Subject Classification

Abstract. The aim of the study was to determine how significance indicators assigned to different Web page elements (internal metadata, title, headings, and main text) influence automated classification. The data collection that was used comprised 1000 Web pages in engineering, to which Engineering Information classes had been manually assigned. The significance indicators were derived using several different methods: (total and partial) precision and recall, semantic distance and multiple regression. It was shown that for best results all the elements have to be included in the classification process. The exact way of combining the significance indicators turned out not to be overly important: using the F1 measure, the best combination of significance indicators yielded no more than 3% higher performance results than the baseline.

1 Introduction

Automated subject classification has been a challenging research issue for several decades now, a major motivation being high costs of manual classification. The interest rapidly grew around 1997, when search engines couldn't do with just full-text retrieval techniques, because the number of available documents grew exponentially. Due to the ever-increasing number of docu-

ments, there is also a danger that recognized objectives of bibliographic systems (finding, collocating, choice, acquisition, navigation) ([19], p. 20-21) would get left behind; automated means could be a solution to preserve them (ibid., p. 30). Automated subject classification of text finds its use in a wide variety of applications, such as: organizing documents into subject categories for topical browsing, which includes grouping search results by subject; topical harvesting; personalized routing of news articles; filtering of unwanted content for Internet browsers; and, many others (see [17], [12]).

A frequent approach to Web-page classification has been a bag-of-words representation of a document, in which all parts of a Web page are considered to be of equal significance. However, unlike other text documents, Web pages have certain characteristics, such as internal metadata, structural information, hyperlinks and anchors, which could serve as potential indicators of subject content. For example, words from title could be more indicative of a page's content than headings. The degree to which different Web page elements are indicative of its content is in this paper referred to as significance indicator.

With the overall purpose of improving our classification algorithm (see section 2.3), the aim was to determine the importance of distinguishing between different parts of a Web page. Significance of four elements was studied: title, headings, metadata, and main text.

The paper is structured as follows: in the second chapter a literature review is given, evaluation issues are discussed and the algorithm used is described (2 Background); in the third chapter data collection as well as methodology for deriving significance indicators are described (3 Methodology); deriving and testing the significance indicators is presented in chapter 4 (4 Significance indicators). The paper ends with conclusions and further research (5 Conclusion).

2 Background

2.1 Related Work

A number of issues related to automated classification of documents and significance of their different parts have been explored in the literature. A. Kolcz, V. Prabakarmurthi, J. Kalita and J. Alspector [14] studied news stories features and found out that initial parts of a story (headline and first two paragraphs) give best results, reflecting the fact that news stories are written so as to capture readers' attention. J. Pierre [16] gained best results in targeted

spidering when using contents of keywords and description metatags as the source of text features, while body text decreased classification accuracy. R. Ghani, S. Slattery & Y. Yang [10] also showed that metadata can be very useful for improving classification accuracy. A. Blum & T. Mitchell [4] compared two approaches, one based on full-text, and one based on anchor words pointing to the target pages, and showed that anchor words alone were slightly less powerful than the full-text alone, and that the combination of the two was best. E. Glover et al. [11] claimed that text in citing documents close to the citation often had greater discriminative and descriptive power than text in target documents. Similarly, A. Attardi, A. Gulli & F. Sebastiani [3] also used information from the context where a URL that refers to that document appears and got encouraging results. J. Fürnkranz [9] used portions of texts from all pages that point to the target page: the anchor text, the headings that structurally precede it, the text of the paragraph in which it occurs, and a set of (automatically extracted) linguistic phrases that capture syntactic role of the anchor text in the paragraph; headings and anchor text proved to be most useful.

On the other hand, R. Ghani, S. Slattery & Y. Yang [10] claim that including words from linked neighborhoods should be done carefully since the neighborhoods could be rather "noisy". Different data collections contain Web pages of various characteristics. If certain characteristics are common to the majority of Web pages in the collection, an appropriate approach taking advantage of those could be applied, but if the Web pages are very heterogeneous, it is difficult to take advantage of any of the Web-specific characteristics (cf. [22], [8], [18]).

2.2 Evaluation Challenge

The problem of deriving the correct interpretation of a document's subject matter has been much discussed in the library science and related literature. It has been reported that different people, whether users or subject indexers, would assign different subject terms or classes to the same document. Studies on inter-indexer and intra-indexer consistency report generally low indexer consistency ([15], p. 99-101). There are two main factors that seem to affect it: 1) higher exhaustivity and specificity of subject indexing both lead to lower consistency (indexers choose the same first term for the major subject of the document, but the consistency decreases as they choose more classes or terms); 2) the bigger the vocabulary, or, the more choices the indexers have, the less likely they will choose the same classes or terms (ibid.).

In this study we start from the assumption that manual classes in our data collection are correct, and compare results of automated classification against them. The classification system used in the study is Engineering Information (Ei), which is rather big (around 800 classes) and deep (five hierarchical levels), allowing many different choices. Without a thorough qualitative analysis of automatically assigned classes we cannot be sure if the classes assigned by the algorithm, which were not manually assigned, are actually wrong.

2.3 Description of the Algorithm

This study is based on an automated classification approach [2] that has been developed within the DESIRE project [6] to produce "All" Engineering [1], an experimental module of the manually created subject gateway Engineering Electronic Library (EELS) [7] (no longer maintained).

The algorithm classifies Web pages into classes of the Ei classification system. Mappings exist between the Ei classes and Ei thesaurus descriptors; both the captions of classes and the descriptors are matched against extracted title, headings, metadata, and main text of a Web page. Each time a match is found, the document is assigned the corresponding class, which is awarded a relevance score, based on which term is matched (single word, phrase, Boolean), the type of class matched (main or optional) (*weight[term]*), and the part of the Web page in which the match is found (*weight[loc]*). A match of a phrase (a number of words in exact order) or a Boolean expression (all terms must be present but in any order) is made more discriminating than a match of a single word; a main class is made more important than an optional class (in the Ei thesaurus, main class (code) is the class to use for the term, while optional class (code) is to be used under certain circumstances). A list of suggested classes and corresponding relevance scores (S) is produced using the following algorithm:

$$S = \sum_{locs} \left(\sum_{terms} \left(freq[loc_j][term_i] * weight[term_i] * weight[loc_j] \right) \right) . \qquad (1)$$

Only classes with scores above a pre-defined cut-off value (cf. section 4.5) are selected as *the* classes for the document. Having experimented with different approaches for stemming and stop-word removal, the best results were gained when an expanded stop-word list was used, and stemming was not applied. For more information on the algorithm, see [2] and [13].

3 Methodology

3.1 Data Collection

The data collection used in the study comprises a selection of Web pages from the EELS subject gateway [7]. EELS Web pages have been selected and classified by librarians for end users of the gateway.

For the study, only pages in English were kept, the reason being that Ei captions and descriptors are in English. Also, some other pages were removed because they contained very little or no text. (The problem of pages containing hardly any text could be dealt with in the future, by propagating the class obtained for their subordinate pages.) The final data collection consisted of 1003 Web pages in the field of engineering.

The data were organized in a relational database. Each document in the database was assigned Ei classes derived from the following elements:

- title (Title);
- headings (Headings);
- metadata (Metadata); and,
- page's main text (Text).

Each class was automatically assigned a score indicating the degree of certainty that it is the correct one. Every document also had manually assigned Ei classes (Manual), against which the automatically assigned classes were compared.

3.2 Methods for Evaluation and Deriving Significance Indicators

Various measures have been used to evaluate different aspects of automated classification performance [21]. Effectiveness, the degree to which correct classification decisions have been made, is often evaluated using performance measures from information retrieval, such as precision and recall, and F1 measure being the harmonic mean of the two. Solutions have also been proposed to measure partial overlap, i.e. the *degree* of agreement between correct and automatically assigned classes (see, for example, [5]).

In this study, three methods have been used for evaluating and deriving the significance of different Web-page elements:

1. total and partial precision, recall, and F1 measures (using macroaveraging);

2. semantic distance; and,
3. multiple regression.

1. The Ei classification system has a solid hierarchical structure, allowing for a rather credible test on partial overlap. Three different levels of overlap were tested: total overlap; partial overlap of the first three digits, e.g. "932.1." and "932.2." are considered the same; and, partial overlap of the first two digits, e.g. "932" and "933" are considered the same. Partial overlap of the first four digits has not been conducted because there were few classes of five-digit length in the data collection.

2. In the literature, different similarity measures have been used for hypermedia navigation and retrieval (see, for example, [20]). Semantic distance, a numerical value representing the difference in meaning between two concepts or terms, is one of them. There are different ways in which to calculate it. For example, the measure of clicking distance in a directory-browsing tree can be used. We used the hierarchical structure of the Ei classification system as the means of obtaining the following (rather arbitrary) measures of semantic distance between any two classes:

- 4, when the classes differ already in the first digit (e.g. 601 vs. 901);
- 2, when the classes differ already in the second digit (e.g. 932 vs. 901);
- 1, when the classes differ in the third digit (e.g. 674.1 vs. 673.1); and
- 0.5, when the classes differ in the fourth digit (e.g. 674.1 vs. 674.2).

Those values reflect how the hierarchical system is structured; e.g. we say that class 6 and class 7 are more distant from each other than classes 63 and 64, which are in turn more distant in meaning than 635.1 and 635.2.

Calculations were conducted using the average distance between manually and automatically assigned classes. For each document, average distances were calculated for each of the four elements, and then the values were averaged for all the documents. When there was more than one manually assigned class per document, the semantic distance was measured between an automatically assigned class and that manually assigned class which was most similar to the automatically assigned one.

3. Multiple regression was used in a rather simplified way: scores assigned based on individual elements of a Web page were taken as independent variables, while the final score represented the dependent variable. The dependent variable was set to either 1000 or 0, corresponding to a correct or an incorrect class respectively.

4 Significance Indicators

4.1 General

Table 1. Distribution of classes in the data collection. First data row shows how many documents have been classified, second row how many classes have been assigned in the whole of the data collection, and the last row how many different individual classes, out of some 800 possible, have been assigned

	Manual	Title	Head-ings	Meta-data	Text
Number of classified doc.	1003	411	391	260	964
In the data collection	1943	827	1504	2227	17089
Different classes	305	174	329	406	675

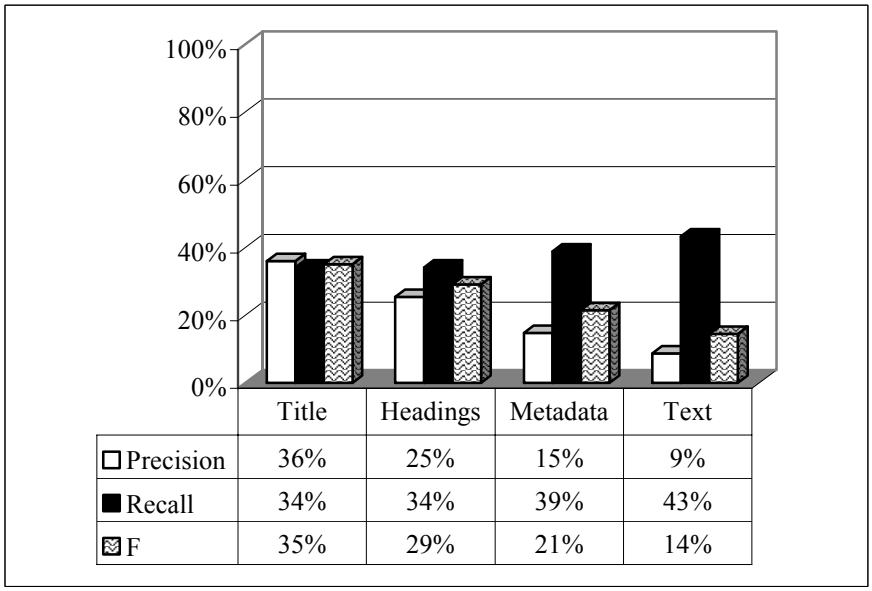
In Table 1 basic classification characteristics and tendencies of our data collection are given. All the documents (1003) have at least one, and no more than six manually assigned classes, the majority having up to three classes. Manual assignment of classes was based on collection-specific classification rules.

Concerning automatically assigned classes based on different parts of a page, not all the pages have classes based on all of them. Classes based on text are assigned to the majority of documents, while those based on metadata to the least number of documents. Based on only title, headings, or metadata, less than 50% of the documents would get classified at all. On the average, per every document there are two manually assigned classes, two classes based on title, four based on headings, nine based on metadata, and some 18 classes based on text.

In the whole collection there are 753 different classes assigned, either manually or automatically. The largest variety comes from the group of classes assigned based on text (675), which is more than twice as many as manually assigned (305).

4.2 Precision and Recall

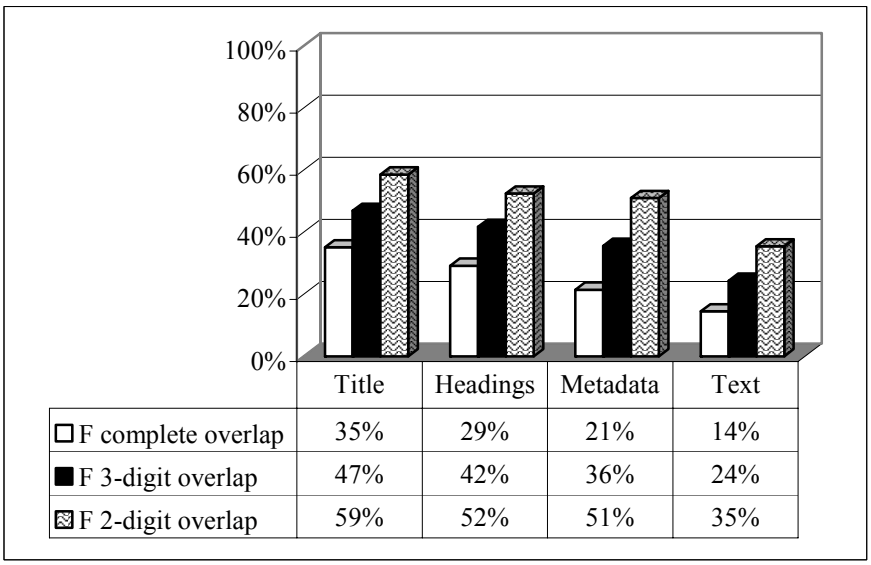
Fig. 1. shows the degree of automated classification accuracy when words are taken solely from the four different parts of the Web page. While title tends to yield best precision, which is 27% more than the worst element (text), text gives the best recall, but only 9% more than the worst element (title). Precision and recall are averaged using the F1 measure, according to which title performs the best (35%), closely followed by headings (29%), metadata (21%) and text (15%).



	Title	Headings	Metadata	Text
Precision	36%	25%	15%	9%
Recall	34%	34%	39%	43%
F	35%	29%	21%	14%

Fig. 1. Precision, recall and F1 measure

Partial Precision and Recall. When testing the algorithm performance for partial overlap (Fig. 2.), precision and recall for all parts of a Web page give much better results (title in 2-digit overlap achieves 59%). The ratio between their performance for both two- and three-digit overlap is the same as for total overlap: title performs the best, followed by headings, metadata and text.



	Title	Headings	Metadata	Text
F complete overlap	35%	29%	21%	14%
F 3-digit overlap	47%	42%	36%	24%
F 2-digit overlap	59%	52%	51%	35%

Fig. 2. F1-measure values for total overlap, 3-digit and 2-digit overlap

4.3 Semantic Distance

Using the semantic distance method, the calculations (Table 2) show that automatically assigned classes are on the average wrong in the third and second digits. Just like precision and recall results for partial overlap (cf. section 4.2), best results (smallest semantic distances) are achieved by title, followed by headings, metadata and text.

Table 2. Semantic distance

	Title	Head-ings	Metadata	Text
Mean distance	1,3	1,7	1,8	2,2

4.4 Deriving Significance Indicators

As we have seen in section 4.1, not every document has all the four elements containing sufficient terms for automated classification. Thus, in order to get documents classified, we need to use a combination of them. How to best

combine them has been experimented with in this section, by applying results gained in evaluation using the F1 measure, semantic distance, and multiple regression.

The symbols used in formulae of this section are:

- S – final score for the automatically assigned class;
- STi – score for the automatically assigned class based on words in Title;
- SH – score for the automatically assigned class based on words in Headings;
- SM – score for the automatically assigned class based on words in Metadata; and,
- STe – score for the automatically assigned class based on words in Text.

The baseline, in which all the elements have equal significance, is represented with the following formula:

$$S = STi + SH + SM + STe . \quad (\mathbf{2})$$

Based on evaluation results, the following co-efficients, representing significance indicators, have been derived (the co-efficients were normalized by reducing the smallest co-efficient to one and by rounding others to integer values):

I. Based on total overlap and F1 measure values:

$$S = 2*STi + 2*SH + SM + STe . \quad (\mathbf{3})$$

These co-efficients have been derived by simply taking the F1 measure values of each of the algorithms (cf. Fig. 1). The same co-efficients have also been derived using partial overlap, the only difference being that the co-efficient for SM was two, both in two- and three-digit overlap.

II. Based on multiple regression, with scores not normalized for the number of words contained in title, headings, metadata, and text:

$$S = 86*STi + 5*SH + 6*SM + STe . \quad (\mathbf{4})$$

III. Based on multiple regression, with scores normalized for the number of words contained in title, headings, metadata, and text:

$$S = STi + SH + SM + 5*STe\ . \qquad (\mathbf{5})$$

IV. On the basis of semantic distance results, the best significance indicator performs less than twice as well as the worst one, so all co-efficients are almost equal, as in (2).

4.5 Evaluation

Defining a Cut-Off. As described in section 2.3, each document is assigned a number of suggested classes and corresponding relevance scores. Only a few classes with best scores, those above a certain cut-off value, are finally selected as *the* classes representing the document.

Different cut-offs, that would give best precision and recall results, were experimented with. Also, the number of documents that would be assigned at least one class, and the number of classes that would be assigned per document, were taken into consideration. Best results were achieved when the final classes selected were those with scores that contained at least 5% of all the scores assigned to all the classes, or, if such a class hadn't existed, the class with the top score was selected. In this case, F1 was 27%, there were about 4000 classes assigned as final, and all documents were classified. This is the cut-off we used in the study.

Results. As seen from Table 3, the evaluation showed that different significance indicators make hardly any difference in terms of classification algorithm performance. Co-efficients in (3) and (5) are similar to the ones in the baseline (2), and, compared to the baseline (2), which performs 23% in F1, normalized multiple regression (5) performs worse by 1%, while the formula based on F1 measure (3) performs the same. The best result was achieved using non-normalized multiple regression (4), which performs by 3% better than the baseline. This formula gives big significance indicator to classes that were assigned based on the title.

Table 3. Results of applying different co-efficients as significance indicators

	Baseline (2)	F1 (3)	Regression (4)	Regression N. (5)
Precision	16%	17%	21%	16%
Recall	39%	39%	35%	38%
F1	23%	23%	26%	22%
Number of pages	1003	1003	1003	1003
Number of classes	5174	5063	4073	5147

5 Conclusion

The aim of this study was to determine the significance of different parts of a Web page for automated classification: title, headings, metadata, and main text. The significance indicators were derived using several different methods: (total and partial) precision and recall, semantic distance, and multiple regression. The study showed that using *all* the structural elements and metadata is necessary since not all of them occur on every page. However, the exact way of combining the significance indicators turned out not to be highly important: the best combination of significance indicators is only 3% better than the baseline.

Reasons for such results need to be further investigated. One could guess that this is due to the fact that the Web pages in our data collection were rather heterogeneous; on the other hand, they were selected by librarians for end users of an operational service, and as such they might indicate what such Web-page collections are like. Apart from heterogeneity, the problem could be that metadata were abused, and that certain tags were misused (e.g. instead of using appropriate tags for making text bold, one used a headings tag, which has the same effect on the screen).

Concerning evaluation of automated classification in general, further research is needed to determine the true value of the classification results. To that purpose information specialists and users could be involved, to compare their judgments as to which classes are correctly assigned. Also, in order to put the evaluation of classification into a broader context, a user study based on different information-seeking tasks would be valuable.

Other related issues of further interest include:

- determining significance of other elements, such as anchor text, location at the beginning of the document versus location at the end, etc.;

- comparing the results with new versions of the Web pages in the collection, e.g. maybe the quality of titles improves with time, and structural tags or metadata get less misused etc.; and,
- experimenting with other Web page collections.

Acknowledgements

The research was funded by ALVIS, an EU Sixth Framework Programme, Information Society Technologies (IST-1-002068-STP), and The Swedish Agency for Innovation Systems (P22504-1 A).

References

1. "All" Engineering Resources on the Internet: A Companion Service to EELS. Available: http://eels.lub.lu.se/ae/ (2003)

2. Ardö, A., Koch, T.: Automatic Classification Applied to the Full-Text Internet Documents in a Robot-Generated Subject Index. In: Online Information 99, Proceedings of the 23rd International Online Information Meeting, London. (1999) 239-246

3. Attardi, G., Gullì, A., Sebastiani, F.: Automatic Web Page Categorization by Link and Context Analysis. In: Hutchison, C., Lanzarone, G. (eds.): Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence. (1999) 105-119

4. Blum, A., Mitchell, T.: Combining Labeled and Unlabeled Data with Co-training. In: Annual Workshop on Computational Learning Theory, Proceedings of the Eleventh Annual Conference on Computational Learning Theory. (1998) 92-100

5. Ceci, M., Malerba, D.: Hierarchical Classification of HTML Documents with WebClassII. In: ECIR. (2003) 57-72

6. DESIRE : Development of a European Service for Information on Research and Education. Available: http://www.desire.org/ (2000)

7. Engineering Electronic Library. Available: http://eels.lub.lu.se/ (2003)

8. Fisher, M., Everson R.: When are Links Useful?: Experiments in Text Classification. In: Proceedings of ECIR-03, 25th European Conference on Information Retrieval, Pisa, IT (2003) 41-56

9. Fürnkranz, J.: Hyperlink Ensembles: A Case Study in Hypertext Classification. Information Fusion 3, 4 (2002) 299-312

10. Ghani, R., Slattery, S., Yang, Y.: Hypertext Categorization Using Hyperlink Patterns and Metadata. In: Proceedings of ICML-01, 18th International Conference on Machine Learning. (2001), 178-185

11. Glover, E.J. et al.: Using Web structure for Classifying and Describing Web Pages. In: Proceedings of the Eleventh International Conference on World Wide Web Honolulu, Hawaii, USA. (2002) 562–569

12. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys 3, 31 (1999) 264-323

13. Koch, T., Ardö, A.: Automatic Classification of Full-Text HTML-Documents from One Specific Subject Area. EU Project DESIRE II D3.6a, Working Paper 2. Available: http://www.it.lth.se/knowlib/DESIRE36a-WP2.html. (2000)

14. Kolcz, A., Prabakarmurthi, V., Kalita, J., and Alspector, J.: Summarization as Feature Selection for Text Categorization. In: Proceedings of the Tenth International Information and Knowledge Management (CIKM-01). (2001) 365-370

15. Olson, H.A., Boll, J.J.: Subject Analysis in Online Catalogs. 2nd ed. Libraries Unlimited, Englewood, Colorado (2001)

16. Pierre, J.: On the Automated Classification of Web sites. In: Linköping Electronic Articles in Computer and Information Science 001 (6) (2001). Available: http://www.ep.liu.se/ea/cis/2001/001/

17. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 1, 34 (2002) 1–47

18. Slattery, S., Craven, M.: Discovering Test Set Regularities in Relational Domains. In: Proceedings of ICML-00, 17th International Conference on Machine Learning. (2000), 895-902

19. Svenonius, E.: The Intellectual Foundations of Information Organization. MIT Press, Cambridge, MA (2000)

20. Tudhope, D., Taylor C.: Navigation via Similarity: Automatic Linking Based on Semantic Closeness. Information Processing and Management, 33(2) (1997) 233-242
21. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval 1/2, 1 (1999) 67-88

22. Yang, Y., Slattery, S., Ghani, R.: A Study of Approaches to Hypertext Categorization. Journal of Intelligent Information Systems. 2/3, 8 (2002) 219-241