Using controlled vocabularies in automated subject classification of textual Web pages, in the context of browsing

Koraljka Golub

Knowledge Discovery and Digital Library Research Group (KnowLib),
Digital Information Systems, Department of Information Technology, Lund University,
P.O. Box 118, 22 100 Lund, Sweden
koraljka.golub@it.lth.se
http://www.it.lth.se/knowlib/

Abstract. Automated subject classification has been a challenging research issue for several decades now. The purpose of this thesis is to determine to what degree controlled vocabularies that have been traditionally used in libraries could be utlised in automated classification of textual Web pages, in the context of browsing. Usefulness of different characteristics of controlled vocabularies for automated classification would be explored, such as captions of classes from classification systems and terms from thesauri and/or subject heading systems. The classification algorithm would be developed based on a research article collection, and tested on Web pages.

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" ([1], p. 259). The term automated subject classification, in the context of this thesis, denotes machine-based organization of related information objects. Certain human intellectual processes are replaced by, for example, statistical and computational linguistics techniques.

Automated subject classification has been a challenging research issue for several decades now. Major motivation has been the high cost of manual classification. The interest has rapidly grown since later 1990s, when search engines couldn't do with just full-text retrieval techniques, because the number of available documents grew exponentially. In the library science community it has been recognized that, due to the ever-increasing number of documents, there is a danger that recognized objectives of bibliographic systems (finding, collocating, choice, acquisition, navigation) ([2], 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, 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 [3], and [4]).

In the narrower focus of this paper is automated classification of textual Web documents into subject categories for browsing. Web documents 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 the 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 ([5]), when one is inexperienced in searching (ibid.), or unfamiliar with the subject in question and its terminology or structure ([10], p.76).

Controlled vocabularies (e.g. classification schemes, thesauri, subject heading systems) have been traditionally 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). They are already used by a number of Web services, especially those providing information services for academic users.

There are three major approaches to automated classification, the biggest being text categorization (coming from the machine-learning community), followed by document clustering (information-retrieval community), and the smallest one, document classification, coming from the library-science community. While the first two approaches use complex algorithms, they hardly utilize controlled vocabularies that are suitable for subject browsing. Library science community research focuses less on algorithms and more on operational systems using controlled vocabularies. 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.

The purpose of this thesis is to determine to what degree controlled vocabularies can be used in automated classification of textual Web pages. The research questions to be dealt with are: to what degree can different elements of classification schemes, and their mappings (to thesauri and/or subject heading systems), improve automated classification; in text categorization, esp. when there is a lack of good-quality training documents for a certain application, what results can be achieved when using the best combination of words from controlled vocabularies instead of words from training documents as class features; and, to what degree can end-users find information resources they are looking for, by browsing classes into which Web pages have been automatically classified (using the approach that gave best results).

The paper is structured as follows: background information on subject browsing, automated classification approaches and controlled vocabularies is given in the following chapter (2 Background); related work is described in the third chapter (3 Related work); and, proposed research with research questions and methodology is given in the last chapter (4 Proposed research).

2 Background

2.1 Subject browsing

Subject browsing in this work refers to seeking for documents through a directory tree of subject classes into which the documents have been classified. Web services offering subject browsing are many, such as those provided by commercial search engines (e.g. [6]), or those provided by quality controlled subject gateways (e.g. [7]; [8]).

Research results have shown that people use subject browsing to a large degree (e.g. [9]) in a number of situations: when users are not looking for a specific item [5], when users are inexperienced in searching (ibid.), when users are unfamiliar with the subject and its structure and terminology (ibid.; [10], p. 76). A. Foskett ([11], p. 13) also claims that users may be browsers, who are looking for something to catch their interest rather than answers to specific questions, and who form the majority of users in public libraries. Browsing also supports serendipity, "the faculty of making happy and unexpected discoveries by accident" (ibid.).

2.2 Controlled vocabularies for subject browsing

Controlled vocabularies have been developed and used in libraries and in indexing and abstracting services, some since the 19th century. These vocabularies can be based on systematic hierarchies of concepts, a variety of relationships defined between the concepts, and they have devices to "control" polysemy, synonymy, and homonymy of the natural language.

There are different types of controlled vocabularies, in this context most interesting being classification schemes, thesauri, and subject heading systems. With the World Wide Web, a new type of controlled vocabulary emerged within computer science and Semantic Web communities: ontologies. Also, directory-style subject browsing found new application in commerical search engines (directories of Web pages).

All these vocabularies have distinct characteristics and are consequently better suited for some classification tasks and 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. Thus classification schemes are better suited for subject browsing than other controlled vocabularies ([12]; [5]; see also [13]). 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 ([10], p. 48).

Subject heading systems and thesauri have traditionally been developed for subject indexing that would describe topics of the document as specifically as possible. Since all these three controlled vocabulary types provide users with different aspects of subject information and different searching functions, their combined usage has been part of the 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 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…" ([10], p. 76).

Although well developed, existing controlled vocabularies need to be improved for the new roles in the electronic environment, such as ([10], p.77-78): 1) improved currency, hospitality for new topics, 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. Some of them are already getting adjusted, such as AGROVOC, the agricultural thesaurus [14], WebDewey, which is Dewey Decimal Classification adapted for the electronic environment, [15], and California Environmental Resources (CERES) thesaurus [16].

2.3 Different automated classification approaches

Text categorization. Text categorization is a machine-learning approach, in which also information retrieval methods are applied. It consists of three main parts. The first part involves manually categorizing 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. Test collections that are used by this community are normally not classified using a classification scheme.

Document clustering. 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.

Document classification. Document classification in this thesis stands for a library science approach. It involves a manually created controlled vocabulary (a classification scheme). Documents are classified into the classes of the used classification scheme. Algorithms tend to be based on string-to-string matching.

Implications for subject browsing. A major difference between the three main approaches to automated classification is in the level of vocabulary control of the used categories. In document classification, controlled vocabularies tend to be well structured for browsing and names used for categories have been carefully chosen. They have devices to control the problems of polysemy, synonymy and homonymy of natural language. In text categorization, the categories' characteristics differ from one test collection to another; they are manually constructed and contain some degree of vocabulary control. However, they neither tend to have the cross-reference structure developed as well as traditional controlled vocabularies, nor is their vocabulary control as thorough in problems of the natural language. Also, often only few categories with one or two hierarchical levels are used, each consequently containing a large, 'unbrowsable' number of documents.

In document clustering, categories are automatically produced, which results in hardly any vocabulary control. Labeling of the clusters is a problem, and relationships between the categories, such as those of equivalence, related-term and hierarchical relationships, are even more difficult to automatically derive ([2], p.168). "Automatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand" [17]. Apart from naming, clusters change as new documents are added to the collection because of the clusters' centroids that are then recalculated. Unstable category names in Web services and digital libraries, for example, are not user-friendly. T. Koch & A. Zettergren [5] suggest that document clustering is better suited for organizing Web search engine results.

2.4 Evaluation challenge

The problem of deriving the correct interpretation of a document's subject matter has been much discussed among library scientists (while 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 ([18], 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.).

3 Related research

Related works include a number of approaches to document pre-processing and indexing parts of automated classification. In all the three approaches to automated classification, most relevant terms from documents need to be selected. Different kinds of terms can be extracted: single words, phrases, stemmed words etc. Based on the contained terms, documents and categories (or classes) are represented as vectors

(text categorization, document clustering), or are compared against terms representing classes (frequent method in document classification). 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. A thorough review of document pre-processing and indexing in text categorization is given by F. Sebastiani ([3], p. 10-18).

The first project aimed at automated classification of Web pages based on a controlled vocabulary was the Nordic WAIS/World Wide Web Project, at Lund University Library and National Technological Library of Denmark [19]. 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 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 (German Harvest Automated Retrieval and Directory) [20] is a robot-generated Web index of Web documents in Germany. 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 built tools for automated subject recognition, using Dewey Decimal Classification (DDC) [33]. 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. In Scorpion, clustering was also used, for refining the result set and for further grouping of documents falling in the same DDC class [21]. Another OCLC project, WordSmith [22], 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.

WWLib (Wolverhampton Web Library) is a manually maintained library catalogue of British Web resources, within which experiments with automated classification were conducted ([23]; [24]). 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 [25] is a robot-generated Web index of about 300000 Web documents, developed within the DESIRE project [26], as an experimental module of manually created subject gateway Engineering Electronic Library (EELS) ([27]; [28]). 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.

4 Proposed research

4.1 Research questions

The purpose of the thesis is to determine to what degree controlled vocabularies could be used in automated classification of textual Web pages. Three major research questions are proposed:

1. To what degree could the following elements of classification schemes, and their mappings (to thesauri and/or subject heading systems), improve automated classification: captions, thesaurus terms, subject heading terms, hierarchical structure, relationships between terms (e.g. related, narrower or broader, is a)?
2. In text categorization, esp. when there is a lack of good-quality training documents for a certain application, what results can be achieved when using the best combination of words from controlled vocabularies instead of words from training documents?
3. To what degree can end-users find information resources they are looking for, by browsing classes into which Web pages have been automatically classified (using the approach that gave best results)?

4.2 Methodology

Test collection. The test collection to be used for developing the classification algorithm should have the following characteristics: a sufficient number of textual documents and metadata describing their content. Each metadata record should contain a manually assigned subject class from a controlled vocabulary.

Controlled vocabularies. Requirements for selecting a controlled vocabulary (probably a classification scheme) would include: a good hierarchical structure, maintenance and up-to-datedness, and mappings to a thesaurus or/and a subject heading system (cf. OCLC's Terminology services [34]).

Variations. A number of parameters will need to be investigated, such as:

1. Which terms to extract from a Web page, e.g. applying a bag-of-words approach or another;
2. Which words to include in a stop-word list;
3. Which weights to assign to extracted terms, e.g. based on *tf***idf* measure;
4. Which cut-off values should be applied.

Evaluation measures. Different measures are used to evaluate different aspects of automated classification performance [30]. Effectiveness, the degree to which correct categorization decision have been made, is often evaluated using performance measures from information retrieval, such as precision and recall; F1 measure is the harmonic mean of the two. Solutions have been proposed to measure partial matching, i.e. the degree of agreement between correct and automatically assigned classes (see, e.g. [31]).

In this thesis, two methods would be used:

1. the standard precision, recall, and F1 measure, based on total and partial matching; and
2. semantic distance.

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. F measure has been defined in the literature as 2*precision*recall/(precision + recall). Good classification schemes have a solid hierarchical structure, thus allowing for a rather credible test on partial matching. Different levels of matching could be tested, e.g.:

1. total match, 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 match, the first three digits, e.g. "932.2.1." and "932.2." have the same first three digits;
3. partial match, the first two digits, e.g. "932" and "933" have the same first two digits.

Semantic distance is here defined as the numerical value representing distance between two classes (cf. [32]). For example, classes "25" and "761.5" are much more semantically distant than classes "243.2" and "243.1" are. Different ways to derive semantic distances would need to be explored.

Evaluation by end-users. Automated classification results would be evaluated by end-users for a number of aspects, such as:

1. How accurate are automatically assigned classes?
2. To what degree can users find needed resources by using the automatically classified resources in the applied browsing structure?

References

1. Chan, L.M.: Cataloging and Classification : an Introduction. McGraw-Hill, New York, 1994
2. Svenonius, E.: The intellectual foundations of information organization. MIT Press, Cambridge, MA (2000)
3. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 1, 34 (2002) 1–47
4. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering : a review. ACM Computing Surveys 3, 31 (1999) 264-323
5. Koch, T., Zettergren, A.-S.: Provide browsing in subject gateways using classification schemes. (1999) http://www.lub.lu.se/desire/handbook/class.html
6. Google Directory. (2005) http://www.google.com/dirhp
7. Resource Discovery Network : RDN. (2004) http://www.rdn.ac.uk
8. Renardus. (2001) http://www.renardus.org
9. Koch, T., Golub, K., Ardö, A.: Users browsing behaviour in a DDC-based web service : a log analysis. Cataloging & Classification Quarterly (2005 *forthcoming*)
10. Schwartz, C.: Sorting out the web : approaches to subject access. Ablex, Westport, CT, (2001)
11. Foskett, A.C.: The Subject Approach to Information. London, Library Assocation Publishing, (1996)
12. Koch, T., Day, M.: The role of classification schemes in Internet resource description and discovery (EU Project DESIRE. Deliverable D3.2.3). (1997)
13. Vizine-Goetz, D.: Using library classification schemes for Internet resources (OCLC Internet Cataloging Project Colloquium position paper). (1996) http://staff.oclc.org/~vizine/Intercat/vizine-goetz.htm
14. Soergel, D. et al.: Reengineering Thesauri for New Applications : The AGROVOC Example. Journal of Digital Information 4, 4 (2004) http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/
15. About DDC : research : a vital part of ongoing development. Available: http://www.oclc.org/dewey/about/research/ (2004)
16. CERES thesaurus effort. (2003) http://ceres.ca.gov/thesaurus/
17. Chen, H., Dumais, S.T.: Bringing Order to the Web : Automatically Categorizing Search Results. Proceedings of CHI-00, ACM International Conference on Human Factors in Computing Systems, Den Haag, NL. New York, ACM Press, (2000)
18. Olson, H.A., Boll, J.J.: Subject analysis in online catalogs. 2nd ed. Libraries Unlimited, Englewood, Colorado (2001)
19. Ardö, A. et al.: Improving resource discovery and retrieval on the Internet : The Nordic WAIS/World Wide Web project summary report. NORDINFO Nytt, 17, 4 (1994) 13-28
20. Möller, G. et al.: Automatic classification of the WWW using the Universal Decimal Classification. Proceedings of the 23rd International Online Information Meeting, London (1999) 231-238

21. Subramanian, S., Shafer, K.E.: Clustering. (1998) http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003409
22. The WordSmith Project. http://www.oclc.org/dewey/about/research/
23. Wallis, J., Burden, P.: Towards a Classification-based Approach to Resource Discovery on the Web. (1995) http://www.scit.wlv.ac.uk/wwlib/position.html
24. Jenkins, C. et al.: Automatic Classification of Web Resources using Java and Dewey Decimal Classification. Computer Networks & Isdn Systems 30, (1998), 646-648
25. "All" Engineering resources on the Internet : a companion service to EELS. (2003) http://eels.lub.lu.se/ae/
26. DESIRE : Development of a European Service for Information on Research and Education. (2000) http://www.desire.org/
27. Engineering Electronic Library. (2003) http://eels.lub.lu.se/
28. 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. (2000) http://www.lub.lu.se/desire/DESIRE36a-WP2.html
29. Prabowo, R. et al.: Ontology-Based Automatic Classification for the Web Pages: Design, Implementation and Evaluation. Proceedings of the 3rd International Conference on Web Information Systems Engineering (2002) 182-191
30. Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval 1/2, 1 (1999) 67-88
31. Ceci, M., Malerba, D.: Hierarchical Classification of HTML Documents with WebClassII. ECIR (2003) 57-72
32. Tudhope, D., Taylor C.: Navigation via Similarity : Automatic Linking Based on Semantic Closeness. Information Processing and Management, 33(2) (1997) 233-242
33. Scorpion. http://www.oclc.org/research/software/scorpion/default.htm
34. OCLC: Terminology services. http://www.oclc.org/research/projects/termservices/