New review of hypermedia and multimedia, Special issue on knowledge organization systems and services, 2006(1)

Automated subject classification of textual Web pages, based on a controlled vocabulary: challenges and recommendations

KORALJKA GOLUB*

KnowLib, Department of Information Technology, Lund University, Sweden

The primary objective of this study was to identify and address problems of applying a controlled vocabulary in automated subject classification of textual Web pages, in the area of engineering. Web pages have special characteristics such as structural information, but are at the same time rather heterogeneous. The classification approach used comprises string-to-string matching between words in a term list extracted form the Ei (Engineering Information) thesaurus and classification scheme, and words in the text to be classified. Based on a sample of 70 Web pages, a number of problems with the term list are identified. Reasons for those problems are discussed and improvements proposed. Methods for implementing the improvements are also specified, suggesting further research.

Keywords: Automated subject classification; Controlled vocabulary; Engineering Information thesaurus and classification scheme

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. Automated classification is also referred to as automated indexing (Moens 2000; Lancaster 2003). In the literature on automated classification and indexing, 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. A major motivation has been high costs of manual subject classification, in terms of time and human resources. The interest has rapidly grown with the advancement of the World Wide Web, on which the number of available documents has been growing exponentially. Due to the ever-increasing number of documents, there is a danger that recognized objectives of bibliographic systems (finding, collocating, choice, acquisition, navigation) (Svenonius 2000, p. 20-21) would get left behind; automated means could be a solution to preserve them (ibid. p.

* Email: Koraljka.Golub@it.lth.se

30). Apart from bibliographic systems, automated classification finds its use in a wide variety of applications, such as grouping search results by subject; topical harvesting; personalized routing of news articles; filtering of unwanted content for Internet browsers; and many others (see Jain, Murty, and Flynn 1999, Sebastiani 2002).

According to K. Golub (2006), one can distinguish between three major approaches to automated classification, the biggest being text categorization (coming from machine-learning community), followed by document clustering (information-retrieval community), and document classification, coming from library science community. While the first two approaches use complex algorithms, they by tradition hardly utilize controlled vocabularies. The library science community research focuses less on algorithms and more on operational systems using controlled vocabularies. The latter approach is more or less based on string-to-string matching of controlled vocabulary terms and text in documents to be classified. Usually weighting schemes are applied with the purpose of indicating degrees to which a term from a document to be classified is significant for the document's topicality. Controlled vocabularies (such as classification schemes, thesauri, subject heading systems) have been traditionally used in libraries, and in indexing and abstracting services, some since the 19th century. They have the devices to control polysemy, synonymy, and homonymy of the natural language, and as such could serve as good-quality structures for subject searching and browsing. Another motivation to apply this approach is to re-use the intellectual effort that has gone into creating such a controlled vocabulary. For further details on the advantages of using pre-existing controlled vocabularies as well as on different approaches to automated classification and indexing see K. Golub (2006), G. Browne (2003a, 2003b), and M.-F. Moens (2000).

String-to-string matching has been explored in linguistics, and controlled vocabularies have been used in automated subject indexing. However, controlled vocabularies largely differ from one another as to their suitability for the task of automated classification or indexing, especially since they have been traditionally designed for *other* tasks. To the author's knowledge, Engineering Information thesaurus and classification scheme (Ei thesaurus 1995) has not been explored in this specific respect by others. In addition, the documents that have been mostly dealt with in these two areas were more traditional document forms, such as research papers, news articles etc., and not Web pages. Web pages 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 misleading, structural tags can be misused, and titles can be without any information significant of the content (e.g. 'Home Page', 'Untitled Document').

This paper is aimed at determining the problems of using controlled vocabularies in automated classification of textual Web pages in the field of engineering, using a string-to-string matching approach based on Ei thesaurus and classification scheme. The study is based on a sample of 70 Web pages and is of a qualitative character.

The paper is laid out as follows: background information with related research and recognized problems are given in the following section; the classification approach used is described in detail in third section; the methodology is given in Section 4; in the last two chapters, the problems are identified and discussed, followed by concluding remarks and recommendations for further research.

2 Related work

A number of projects and studies have been conducted for the purpose of classifying Web pages, using classification schemes. In the Nordic WAIS/World Wide Web Project (Ardö *et al.* 1994), World Wide Web documents and WAIS (Wide Area Information Server) databases were being automatically classified, using Universal Decimal Classification (UDC). 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; by comparing the extracted words with UDC's vocabulary a list of suggested classifications was generated; the words were weighted based on which part of the description they belonged to. They report that 10% of the total 660 databases don't get any classification at all, which is mostly because there are no significant keywords in the descriptions, and suggest that they would extend their UDC vocabulary, which would help solve this problem.

A later project GERHARD (German Harvest Automated Retrieval and Directory) was aimed at creating a robot-generated Web index of Web documents in Germany (Möller *et al.* 1999). The Web index was based on a multilingual version of UDC in English, German and French. GERHARD's approach involved advanced linguistic analysis: 1) processing captions[**], which included stop-words removal, morphological analysis of each word and its reduction to the stem; 2) removing prefixes and stop-words from Web pages. The words and phrases were then extracted and matched against the captions. The resulting set of UDC notations was ranked and weighted statistically according to frequencies and to the structure of the document.

A major Web page classification project was OCLC's (Online Computer Library Center) Scorpion, within which tools were built for automated subject recognition, using Dewey Decimal Classification (DDC). The basic idea was to treat a document to be indexed as a query against a DDC knowledge base. The results of the 'search' were treated as subjects of the document. Scorpion also used clustering, to refine the result set and to further group documents falling in the same DDC class (Subramanian, Shafer 1998). Another OCLC project, WordSmith (The WordSmith Project), provided support for automated classification. The software developed used a variety of computational linguistics methods 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 of the complete text of the raw document. However, it showed that there were no significant differences.

Other examples of automated classification of Web pages using controlled vocabularies include WWWLib (Wolverhampton Web Library), a manually maintained library catalogue of British Web resources, within which experiments with automated classification were conducted (Wallis and Burden 1995, Jenkins *et al.* 1998). The earlier study is based on matching of words from the document against words in DDC and reports 34% correctly classified documents. The latter study takes advantage of DDC's hierarchy: words are extracted from the document and assigned weights; they are first matched at the top level, proceeding further down the hierarchy until a significant match is found with a leaf node. Also, in a study by R. Prabowo et al. (2002) ontologies were built with which to classify Web pages. The ontologies were based on DDC and Library of Congress Classification (LCC). In the process of Web page classification, a feed-forward neural network was used to assist the

[**] A caption is a class number expressed in words; e.g. in UDC, 'telescopes' is the caption for class '520.2'.

classifier in measuring the similarity between the Web page and a class representative. Term weighting was based on position on a Web page, and the number of term occurrences. They claim that their approach results in improved classification accuracy, but also point to the problem of ontology incompleteness in the process of automated classification.

Related work also includes automated classification of documents other than Web pages, using controlled vocabularies such as MeSH (Medical Subject Headings) (NLM's Indexing Initiative; Roberts and Souter 2000) or INSPEC thesaurus (Plaunt and Norgard 1997; BINDEX 2001). Lexis-Nexis has developed an approach called SmartIndexing®, based on a controlled list of terms and their profiles, created by indexers, while assignment of those terms is done automatically (Tenopir 1999). In a study similar to ours, but based on medical documents and applying International Code of Disease (ICD) classification scheme (Ribeiro-Neto, Laender, and Lima 2001), five different cases of the classification algorithm failures were discovered: no class found, due to the fact that the ICD alphabetical index is incomplete; a class was found but didn't correspond to the manually assigned class because the specialist didn't assign an appropriate class; names for narrower concepts didn't exist in the alphabetical index; the presence of a human expert is required since specialist knowledge is needed to deduce the class; and, the class assigned by the algorithm was wrong.

3 Approach

3.1 Algorithm

The study is based on an automated classification approach (Koch and Ardö 2000) that has been developed within the DESIRE project (DESIRE) to produce 'All' Engineering ('All' Engineering resources on the Internet 2003), an experimental module of the manually created subject gateway Engineering Electronic Library (EELS) (Engineering Electronic Library) (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. A list of suggested classes and corresponding relevance scores (S) is produced using the following algorithm:

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

Term weight ($weight[term_i]$) is taken from the term list (see 3.2). Location weight ($weight[loc_j]$) is defined for locations like title, metadata, HTML headings, and main text. The applied formula was 86*scoreTitle, 5*scoreHeadings, 6*scoreMetadata, 1* scoreText, as determined in K. Golub and A. Ardö 2005. Frequency ($freq[loc_j][term_i]$) is the number of times $term_i$ occurs in the text of location loc_j.

Only classes with scores above a pre-defined cut-off value are selected as *the* classes for the document: best results are achieved when the final classes selected are those with scores that contain at least 5% of the sum of all the scores assigned in total, or, if such a class doesn't exist, the class with the top score is selected. According to

the policies for the collection, on the average 3 classes per document are automatically assigned.

Having experimented with different approaches for stemming and stop-word removal, best results were gained when an expanded stop-word list was used, and when stemming was not applied – stemming was shown to improve recall at the expense of precision (Koch and Ardö 2000, Ch.5).

Precision, recall and F1 measure were used as standard evaluation measures (Moens 2001, p.104-105). Precision is the ratio of correct automatic assignments divided by the total number of correct automatic assignments. Recall is the ratio of correct automatic assignments divided by the total number of automatic assignments. The F1 measure is a combination of precision and recall:

F1(precision, recall) = 2*precision*recall / (precision + recall) .

By comparing automatically assigned classes to manually assigned ones at all the five levels of specificity (Ei has five hierarchical levels), the F1 measure was 0,26, whereas if comparison was done by reducing all the classes to the first two hierarchical levels, F1 was 0,59 (K. Golub and A. Ardö 2005). Also, an additional evaluation was performed, in which a subject expert evaluated both the automatically and manually assigned classes of a random sample of 109 Web pages. Based on this type of evaluation, the automated approach gained the F1 of 0,66.

3.2 Term list

The term list used in our approach is based on the Ei (Engineering Information) thesaurus and classification scheme (Ei thesaurus 1995). Here is an extract from the classification scheme:

4 Civil Engineering
…
44 Water and Waterworks Engineering
441 Dams and Reservoirs
…
445 Water Treatment
445.1 Water Treatment Techniques
445.1.1 Potable Water Treatment Techniques
…

It contains the following types of terms and relationships: descriptors and their synonyms, related terms, broader and narrower terms, and scope notes. Thesaurus descriptors are mapped to classification codes. The classification codes are organized into six categories which are divided into 38 subjects, which are further subdivided into 182 specific subject areas. These are yet further subdivided, resulting in over 800 individual classes in a five-level hierarchy. There are some 20 000 terms in the thesaurus, with an average of 11 terms assigned per class.

The term list contains class captions, descriptors and synonyms, and their mappings to Ei classes. It is organized as a list of triplets: term (single word, Boolean term, phrase), class to which it maps, and weight. Boolean terms consist of words that must all be present but in any order or in any distance from each other. The Boolean terms are not explicitly part of the Ei thesaurus, so they had to be created in a pre-processing step. They are considered to be those terms from the Ei thesaurus, which contain the following strings: ‘and’ (word *and*), ‘vs.’ (short for *versus*), ‘,’ (comma), ‘;’ (semi-colon), ‘(‘ (bracket), ‘:’ (colon), and ‘--’ (double dash).

Concerning weighting, a main class is made more important than an optional class: in the Ei thesaurus, main class is the class to use for the term, while optional class is to be used under certain circumstances. Phrases are assigned the highest weights (40 for main class, 20 for optional class), since they normally are most discriminatory. Boolean AND-expressions are the next best (15 for main class, 10 for optional class). Single words can be too general and/or have several meanings or uses that make them less specific and should thus be assigned a small weight. In the first run of the experiments, they were assigned the same weights as Boolean entries (15 for main class, 10 for optional class).

Upper-case words from the Ei are left in upper case in the term list, assuming that they are acronyms. All other words containing at least one lower-case letter are converted into lower case.

Here is an excerpt from the Ei thesaurus, based on which the excerpt from the term list (further below) was created:

TM Active solar buildings
MC 402
MC 643.1
MC 657.1

TM Catalyst activity
UF Catalysts--Activity
MC 803
MC 804
OC 802.2

TM stands for the descriptor, and UF for synonym; MC represents the main class, and OC an optional class. Below is an excerpt from the term list, as based on those two examples:

40: active solar buildings=657.1, 643.1, 402,
15: activity @and catalysts=803, 804,
10: activity @and catalysts=802.2,
40: catalyst activity=803, 804,
20: catalyst activity=802.2,

3.3 Data collection

The data collection used in the study comprises a sample of 1 000 Web pages from the EELS subject gateway (Engineering Electronic Library). EELS Web pages have been selected and classified by librarians for end users of the gateway. Using the described algorithm (see 3.1), each class is automatically assigned a score indicating the degree of certainty that it is the correct one. Every document also has manually assigned Ei classes, against which the automatically assigned classes were compared.

4 Methodology

To the purpose of identifying problems of the described approach to automated classification, a qualitative analysis was applied to selected 70 Web pages. These were Web pages from the data collection (see 3.3), which had all the automatically assigned classes wrong, at least in comparison to the (pre-existing) intellectually assigned classes.

Each Web page in the sample was thoroughly examined. The entire text of the Web page was read in detail, including title, headings, metadata and the rest of the Web page's content. *All* automatically assigned classes were looked at and compared against the intellectually assigned ones, even those below the cut-off value, which were not selected as *the* classes for the document (cf. 3.1). The nature of problem for

each automatically assigned class that was not intellectually assigned was specified as being one of the following:

- class not found at all (112 instances in the sample);
- class found but below threshold (71 instances in the sample);
- wrong automatically assigned class (148 instances in the sample); and,
- automatically assigned class that is not really wrong (36 instances in the sample).

For each class it was determined why it was automatically assigned, by looking at terms found on a Web page and comparing them against terms in the term list designating corresponding classes.

5 Problems identified

5.1 Class not found at all

In the sample, two major types of disagreement have been identified:

a) Intellectually assigned and automatically assigned classes are of equal length, i.e. they are at the same hierarchical level of specificity: first, second, or third. Several cases were found where classes overlapped in the first three digits, i.e. they were both at the third hierarchical level (e.g. 921.5 vs. 921.3), and several more overlapped in the first two digits (e.g. 402 vs. 409) or in the first digit only (e.g. 901.4 vs. 912.3). Also, in several instances, intellectually and automatically assigned classes did not overlap even in the first digit (e.g. 901.3 and 723.3).
b) Intellectually assigned and automatically assigned classes are not of equal length, i.e. they are at different levels of specificity. In other words, automatically assigned classes are either broader than manually assigned ones (e.g. 723 vs. 723.5), or narrower (e.g. 723.1.1 vs. 72). Also, in several instances, intellectually and automatically assigned classes did not overlap not even in the first digit (e.g. 655.1 and 731).

These classes were not found because the words in the term list designating the classes were not found in the text of the Web page to be classified. The same problem has been discovered in automated classification of medical documents (Ribeiro-Neto, Laender and Lima 2001, p.398).

5.1.1 Recommendations.

An ideal approach would be to re-design the Ei thesaurus and classification scheme by adding more synonyms, as well as introducing new concepts. Synonyms could be introduced based on an intellectually produced synonym list such as WordNet (WordNet), although the right sense for each term would also need to be selected manually. Another alternative would be to automatically extract additional terms for each class, from documents known to belong to the class.

To provide for different word forms and different ordering of words in a term, regular expressions could be used, although they need to be manually introduced to

the term list. An automated alternative would be to apply natural language processing tools and methods.

5.1.2 Example demonstrating the problem and possible solutions.

URL: http://www.iiasa.ac.at/ (as downloaded at the time of the classification process)
Words from title: IIASA home page
Words from metadata: /
Words from headings: international institute for applied systems analysis
Words from the main text: IIASA LOGO welcome to the international institute for applied systems analysis IIASA is a non-governmental research organization located in austria international teams of experts from various disciplines conduct scientific studies on environmental economic technological and social issues in the context of human dimensions of global change the institute is sponsored by 17 national member organizations in north america europe and asia for more details home - what's new - general info - research publications - options - world - web map - search international institute for applied systems analysis. A-2361 laxenburg austria phone 43-2236-807-0 fax 43-2236-71313 web support for optimal viewing of this site use at least netscape navigator 3 and adobe acrobat reader 3
One of the manually assigned classes: 901.4, which stands for 'Impact of Technology on Society'
The automatically assigned class that is closest to the above manually assigned class: 912.3, which stands for 'Operations Research' and belongs to 912 class for 'Industrial Engineering and Management'
Term list for 901.4:
20: behavioral science computing=911.2, 912.2, 901.4, 461.4,
20: data processing @and social @and behavioral sciences applications=911.2, 912.2, 901.4, 461.4,
20: economic @and social effects=911.2, 901.4,
20: economic effects=911.2, 901.4,
40: engineering @and social aspects=901.4,
40: impact of technology on society=901.4,
20: public risks=901.4,
20: risk studies @and public risks=901.4,
20: robots @and industrial @and socioeconomic aspects=911.2, 901.4,
40: social aspects=901.4,
20: social effects=911.2, 901.4,
20: social sciences computing=911.2, 912.2, 901.4, 461.4,
20: social sciences=911.2, 912.2, 901.4, 461.4,
20: socioeconomic effects=911.2, 901.4,
40: sociological aspects=901.4,
20: sociological effects=911.2, 901.4,
20: technological forecasting=901.4,
20: technology @and economic @and sociological effects=911.2, 901.4,
20: technology transfer=911.2, 901.4,
Term list for 912.3:
10: PERT=912.3,
20: bioengineering @and systems science=731.1, 461.1, 912.3,
20: complex systems=731.1, 461.1, 912.3,
20: composite systems=731.1, 461.1, 912.3,
20: cost effectiveness=912.2, 912.3,
20: data processing @and PERT=912.3,
20: interconnected systems=731.1, 461.1, 912.3,
20: large scale systems=731.1, 461.1, 912.3,
40: operations research=912.3,
20: program evaluation @and review technique=912.3,
20: resource allocation=912.3,
40: system analysis=912.3,
40: system design=912.3,
20: system science=731.1, 461.1, 912.3,
20: system theory=461.1, 912.3, 731.4,
40: systems analysis=912.3,
40: systems design=912.3,

20: systems science @and cybernetics @and large scale systems=731.1, 461.1, 912.3,
20: systems science @and cybernetics @and system theory=461.1, 912.3, 731.4,
20: systems science @and cybernetics=731.1, 461.1, 912.3,
20: systems science=731.1, 461.1, 912.3,
20: systems theory=461.1, 912.3, 731.4,

Results of automated classification: Based on the term list, class 912.3 was assigned since one of the terms designating it, 'systems analysis', was found in the headings and main text of the Web page. The matching instances are marked by grey rectangles. The 912.3 was automatically assigned a score of 280, according to the formula given in 3.1:

$$Score_{912.3} = 1_{frequency} * 5_{headings} * 40_{termtype} + 2_{frequency} * 1_{plaintext} * 40_{termtype} = 280$$

Examples of suggested improvement: Enriching the term list for 901.4 with synonyms, such as the following ones, would be beneficial:

environmental issues
economic issues
technological issues
technological change
social issues
systems analysis

In a similar text, one could find different word forms, such as noun, verb and adjectival forms, singular or plural etc. For example, the following could be introduced to the 901.4 term list:

social science
technology @ forecasting
sociology @ effects
sociology @ effect

Also, introducing different ordering should be provided for, e.g. for 'impact of technology on society' we could have:

technology @ impact @ society
technology @ impact
technology @ society

5.2 Class found but below threshold

The main reason for not assigning correct classes that were discovered automatically has to do with weighting and cut-off values. This is because only classes with scores above a pre-defined cut-off value are selected as *the* classes for the document: the final classes selected are those with scores that contain at least 5% of the sum of all the scores assigned in total, or, if such a class doesn't exist, the class with the top score is selected (cf. 3.1). Another reason could be that the classification algorithm is made to always pick the most specific class as the final one, which is in accordance with the given policy for intellectual classification.

5.2.1 Example demonstrating the problem.

URL: http://www.luth.se/depts/mt/hallf/index2.html (as downloaded at the time of the classification process)

Words from title: solid mechanics hållfasthetslära luleå univ of technology Sweden

Words from metadata: /

Words from headings: /

Words from the main text (excerpt): staff research areas current projects equipment what is solid mechanis under graduate education… material is called solid rather than fluid if it can also fracture mechanics computational computer numercal simulations upport a substantial shearing force over the time scale of some natural process or technological application of interest shearing forces are directed parallel rather than perpendicular to the material surface on which they act the force per unit of area is called shear stress for example consider a vertical metal rod that is fixed to a support at its upper end…

Two of the manually assigned classes: 421, which stands for 'Strength of Building Materials; Mechanical Properties', and 422, which stands for 'Strength of Building Materials; Test Equipment and Methods'

Automatically derived classes selected as the classes for the document: 931.1, which stands for 'Mechanics', 901.2, which stands for 'Education', 901.3, which stands for 'Engineering Research' and

901 for 'Engineering Profession'. These selected classes were the ones that had a score containing at least 5% of the sum of all the scores assigned in total; as given below, the sum of all the scores of all the automatically assigned classes was 11775.
All the automatically derived classes for the document: 38 different classes were automatically derived and ranked (score is in the brackets):
931.1 (3795), 901.2 (1935), 901.3 (1845), 901 (1815), 421 (525), 933.2 (150), 481.1.2 (120), 933.1 (120), 804.2 (105), 481.1 (105), 933 (90), 604.1 (90), 641.1 (90), 657.2 (75), 931.3 (60), 535.1 (60), 804 (60), 931.2 (60), 818.1 (60), 741.1 (60), 412 (45), 444 (45), 422 (45), 657 (45), 408.1 (45), 802.3 (45), 414 (45), 545.3 (30), 483.1 (30), 461.2 (30), 812.3 (30), 531.1 (15), 903.2 (15), 801.4 (15), 932.2 (15), 482.2 (15), 505 (15), 631 (15).

5.2.2 Recommendations.

Experiment with different heuristics for weights and cut-offs. For example, weights for terms could be determined based on document vs. collection frequencies, or based on how many documents in which the term occurs are actually relevant vs. the number of documents where it occurs but is not relevant for the document (cf. Salton and Buckley 1988). Statistical methods such as multiple regression could also be used, to derive appropriate weights automatically (cf. Golub, Ardö 2005, p.372).

A different solution could be to include all the automatically found classes, and assign them the automatically derived weights. Weights indicating term importance for a certain document have also been attributed by human indexers performing intellectual (manual) indexing (Moens 2000, p.58).

5.3 Wrong automatically assigned class

Based on the sample, four different sub-problems have been identified:

a) words recognized as homonyms or distant synonyms, e.g.:

- 'association' - a Web page on 'international association of drilling contractors' gets wrongly classified as belonging to class 'chemical reactions', because the word 'association' in the term list is mapped to that class;
- 'paper' – a Web page of the 'journal of the electrochemical society' is wrongly classified as belonging to class 'pulp and paper' because of the word 'paper', which is on the Web page found in the context of research papers published in the journal, but in the term list it is mapped to the class 'pulp and paper';
- 'architecture' and 'facilities' – a Web page on computers is wrongly classified as belonging to class 'buildings and towers' because the word 'architecture' referring to computer architecture, is found on the Web page but in the term list is mapped to that class; a Web page on 'labs and facilities' is also wrongly classified as 'buildings and towers' because the word 'facilities' in the term list mapped to that class;
- 'information technology' – a Web page on computer information center is wrongly classified as belonging to class 'information science' because the term 'information technology' found on the Web page is in the term list mapped to that class;
- 'systems analysis' – a Web page about an institute studying technological impact on society is wrongly classified as belonging to class 'control systems' in control engineering, because the term 'systems analysis' is mapped to that class;
- 'safety' – a Web page about the world wide web virtual library safety is wrongly classified as belonging to class 'accidents and accident prevention' in engineering, because the word 'safety' in the term list is mapped to that class;

- ‘hardware’ – a Web page containing ‘bibliographies on software/hardware engineering and formal methods’ is wrongly classified as belonging to class ‘small tools and hardware’ because the word ‘hardware’ is mapped to that class;

b) word found on a Web page is there because it is an instance of what it represents, and it is not *about* such instances, e.g.:

- a Web page containing bibliographies or allowing access to databases on computer science is classified as ‘information science’ or as ‘database systems’, instead of being classified as ‘computer science’;
- a Web page on SQL standard or a Web page on a classification system get wrongly classified as ‘codes and standards’ in engineering;
- a Web page that is an information service for artificial intelligence gets classified as ‘information services’;
- a Web page on online tutorials and e-learning programs for technical fields gets wrongly classified as a Web page on ‘education’;

c) too distant term–class mappings, including cases when one term in the term list is mapped to several different classes, e.g.:

- the word ‘bibliographies’ is mapped to ‘information science’;
- terms ‘policy’, ‘technology’, and ‘public policy’ are all mapped to ‘engineering profession’;
- the word ‘air’ is mapped to ‘chemical products generally’;
- the word ‘textbooks’ is mapped to ‘education’ and ‘information dissemination’;
- the term ‘social sciences’ is mapped to ‘human engineering’, ‘industrial economics’, ‘management’ in industrial engineering, and ‘impact of technology on society’;
- the word ‘cryptography’ is mapped to ‘telephone and other line communications’, ‘electronic equipment, radar, radio and television’, ‘electro-optical communication’, and ‘computer software, data handling and applications’;

d) words mentioned on the Web page have little to do with the Web page’s aboutness, e.g. an institution’s Web page gets wrongly classified as ‘facsimile systems and technology’, because among their contact information, there is also a fax number, and the word ‘fax’ is mapped to that class.

5.3.1 Recommendations.

Word-sense disambiguation in context is needed. Homonym and polysem resolution could be improved by introducing the rule in the algorithm that, before making a final decision whether to assign a certain class, it checks if similar classes (e.g. within the same top-level class) have also been assigned to that document. For example, if class 811.1 for ‘Pulp and Paper’ gets assigned because the word ‘paper’ is found (in the term list: ‘15: paper=811.1’), the algorithm would check if other classes starting with ‘811’ have also been automatically assigned:

811 :: Cellulose, Paper and Wood Products
811.1.1 :: Papermaking Processes
811.1.2 :: Papermaking Equipment
811.2 :: Wood and Wood Products
811.3 :: Cellulose and Derivatives

Another approach would be to classify also sub-pages of the Web page being classified to confirm the classification of the main page. For example, if class 811.1 for 'Pulp and Paper' gets assigned to a Web page, the algorithm could check if the Web pages to which this page links are assigned other classes starting with '811.' This could at the same time, also solve the problem of Web pages containing hardly any text.

It is also possible to provide context in the term list itself, by applying Boolean operators to create new, context-enriched terms, e.g. by adding a broader term or class caption to single words and homonyms. For example, the term item '15: association=802.2,' could be replaced by combining the single word 'association' with words from classes at level 800, for 'Chemical Engineering':

40: association @and chemical engineering=505.1

It might help if single words are given lower weight as well.

A different approach would be to enrich the context by introducing synonyms of the correct senses from WordNet (WordNet), which, again, needs to be done manually. Yet another way to proceed would be to determine in which context in the document collection homonyms mostly occur; if a homonym most often occurs in a non-Ei context, put it on the stop list, or use a large negative weight in the term list. Also individual rules could be introduced, such as, the class 'facsimile systems and technology' could be assigned when there aren't any numbers following the word 'fax'. Ideally, there would also be an expert going through the list and removing those entries that are too distant in meaning from the class they are to designate.

Another problem is that the classification algorithm assigns a class based on a Boolean term, no matter how far from each other elements of the Boolean term are on a Web page. A proximity measure could be introduced in the algorithm, defining that, for example, two words connected by a Boolean AND should not have more than seven words between them.

5.4 Automatically assigned class that is not really wrong

In the sample, a number of classes were found that were not really wrong, but were not intellectually assigned. For example, a Web page on chemistry and biochemistry is automatically assigned a class for 'biology'; or, a Web page on 'mechanical engineering, plant and power' and 'electric transmission and distribution' is assigned 'nuclear reactors', which is not really wrong because 'nuclear power' is also what it is about. This could have to do with the subject indexing policy, such as exhaustivity (cf. Moens 2000, p.72).

5.4.1 Recommendations.

Further research is needed to determine to what degree and in which contexts classes that are automatically assigned are to be discarded as incorrect or could actually be useful to end-users. Methodology for such user-based studies needs to be developed.

6 Concluding remarks

In the study, several different problems of string-to-string matching approach to automated classification were identified and discussed. The focus of the study was a collection of Web pages in the field of engineering. Web pages present a special

challenge: because of their heterogeneity the same principle (e.g. words from headings are more important than main text) is not applicable to all the Web pages of a collection. For example, utilizing information from headings on all Web pages might not give improved results, since headings are sometimes used simply instead of using bold or a bigger font size.

The matching was based on a term list derived from the Ei thesaurus and classification scheme, which contains a number of different types of terms and relationships, and thesaurus descriptors are mapped to classification codes. As a result of these elaborate relationships, on average 11 terms get assigned per class. This allows for relatively good classification results by using only the simple string-to-string matching.

However, a number of weaknesses of the described approach were identified, and ways to deal with those were proposed for further research. These include enriching the term list with synonyms and different word forms, adjusting the term weights and cut-off values and word-sense disambiguation. In our further research the plan is to implement automated methods. On the other hand, the suggested manual methods (e.g. adding synonyms) would, at the same time, improve Ei's original function, that of enhancing retrieval. Having this purpose in mind, manually enriching a controlled vocabulary for automated classification or indexing would not necessarily create additional costs.

Acknowledgements

The author would like to thank Anders Ardö, reviewers and editors whose suggestions helped improve the paper.

The research was funded by ALVIS, an EU Sixth Framework Programme, Information Society Technologies (IST-1-002068-STP).

References

'All' Engineering resources on the Internet : a companion service to EELS. Available online at: http://eels.lub.lu.se/ae/ (accessed 13 January 2006).

Ardö, A., et al., Improving resource discovery and retrieval on the Internet: the Nordic WAIS/World Wide Web project summary report. *NORDINFO Nytt*, 1994, **17**(4), 13-28.

Browne, G. Automatic categorisation. Part 1: Principles of classification. Online Currents, 2003a, **18**(1), 17-22.

Browne, G (2003b). Automatic categorisation. Part 2: Technology. Online Currents, 2003b, **18**(2), 7-11.

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

DESIRE: Development of a European Service for Information on Research and Education. Available online at: http://www.desire.org/ (accessed 22 December 2004)

Ei thesaurus, edited by J. Milstead, Engineering Infromation, Castle Point on the Hudson Hoboken, 1995. 2nd ed.

Engineering Electronic Library. Available online at: http://eels.lub.lu.se/ (accessed 13 January 2006).

Figuerola, C.G., Rodriguez, A.F.Z., and Berrocal, J.L.A. Automatic vs manual categorisation of documents in Spanish. *Journal of Documentation*, 2001, **57**(6), 763-773.

Golub, K., Automated subject classification of textual Web documents. Accepted for publication in *Journal of Documentation*, 2006, **62**. Available online at: http://www.it.lth.se/koraljka/Lund/publ/AC-JDoc.pdf (accessed 13 January 2006).

Golub, K., and Ardö, A., Importance of HTML structural elements and metadata in automated subject classification, in *Research and advanced technology for digital libraries, Proceedings of ECDL 2005 – the 9th European conference on research and advanced technology for digital libraries, Vienna, Austria, 18-23 September*, 2005, pp. 368-378.

Hjørland, B., Lifeboat for knowledge organization: indexing theory. Available online at: http://www.db.dk/bh/Lifeboat_KO/CONCEPTS/indexing_theory.htm (accessed 13 January 2006).

BINDEX, Available online at: http://web.archive.org/web/20050323163031/http://www.hltcentral.org/projects/print.php?acronym=BINDEX (accessed 18 April 2006).

International Organization for Standardization. Documentation – Methods for examining documents, determining their subjects, and selecting index terms: ISO 5963-1985. Geneva, Switzerland: International Organization for Standardization.

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

Jenkins, C. et al., Automatic classification of Web resources using Java and Dewey Decimal Classification. *Computer Networks & Isdn Systems*, 1998, **30**, 646-648.

Koch, T., and Ardö, A., Automatic classification of full-text HTML-documents from one specific subject area. EU Project DESIRE II D3.6a, Working Paper 2. 2000. Available online at: http://www.lub.lu.se/desire/DESIRE36a-WP2.html (accessed 13 January 2006).

Lancaster, F.W., Indexing and abstracting in theory and practice. 3rd ed. London: Facet, 2003.

Moens, M.-F. Automatic indexing and abstracting of document texts. Boston: Kluwer, 2000.

Möller, G. et al., Automatic classification of the WWW using the Universal Decimal Classification, in *Proceedings of the 23rd International online information meeting*, 1999, pp. 231-238.

NLM's Indexing Initiative. Available online at: http://ii.nlm.nih.gov/ (accessed 13 January 2006).

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

Plaunt, C., and Norgard, B.A., An association-based method for automatic indexing with controlled vocabulary. *Journal of the American Society for Information Science*, 1998, **49**(10), 887-902.

Prabowo, R. et al., Ontology-based automatic classification for the Web pages: design, implementation and evaluation, in *Proceedings of the 3rd International conference on Web information systems engineering*, 2002, pp. 182-191.

Ribeiro-Neto, B, Laender, A.H.F., and de Lima, L.R.S. An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science and Technology*, 2001, **52**(5), 391-401.

Roberts, D., and Souter, C., The automation of controlled vocabulary subject indexing of medical journal articles, in *Aslib Proceedings*, 2000, 52(10), pp. 384-401.

Salton, G., and Buckley, C. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 1988, **24**(5), 513–523. Reprinted in: K. Sparck Jones, P. Willett (Eds.), Readings in information retrieval (pp. 323–328). San Francisco: Morgan Kaufman, 1997.

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

Subramanian, S., and Shafer, K.E. Clustering. 1998. Available online at: http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003409 (accessed 13 January 2006).

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

Tenopir, C., Human or automated, indexing is important. *Library Journal*, 1999, **124** (18) 34–38.

Wallis, J., and Burden, P., Towards a classification-based approach to resource discovery on the Web. 1995.

Available online at: http://www.scit.wlv.ac.uk/wwlib/position.html (accessed 13 January 2006).

WordSmith Project. Available online at: http://www.oclc.org/dewey/about/research/ (accessed 13 January 2006).

WordNet. Available online at: http://wordnet.princeton.edu/ (accessed 13 January 2006).