Automated Classification of Web Pages in Hierarchical Browsing

Abstract

Purpose

The purpose of the study was twofold: to investigate whether it is meaningful to use the Ei classification scheme for browsing, and then, if proven useful, to investigate the performance of an automated classification algorithm based on the Ei classification scheme.

Design/methodology/approach

A user study was conducted in which users solved four controlled searching tasks. The users browsed the Ei classification scheme in order to examine the suitability of the classification systems for browsing. The classification algorithm was evaluated by the users who judged the correctness of the automatically assigned classes.

Findings

The study showed that the Ei classification scheme is suited for browsing. Automatically assigned classes were on average partly correct, with some classes working better than others. Success of browsing showed to be correlated and dependent on classification correctness.

Research limitations/implications

Further research should address problems of disparate evaluations of one and the same web page. Additional reasons behind browsing failures in the Ei classification scheme also need further investigation.

Practical implications (if applicable)

Improvements for browsing were identified:

- Describing class captions and/or listing their subclasses from start;
- Allowing for searching for words from class captions with synonym search (easily provided for Ei since the classes are mapped to thesauri terms);
- When searching for class captions, returning the hierarchical tree expanded around the class in which caption the search term is found.

The need for improvements of classification schemes was also indicated.

Originality/value

A user-based evaluation of automated subject classification in the context of browsing has not been conducted before; hence the study also presents new findings concerning methodology .

Keywords: automated subject classification, hierarchical browsing, Ei classification scheme and thesaurus, evaluation, user studies

Classification: Research paper

1 Introduction

Automated subject classification research began with the availability of electronic text in the early 1950s. The first idea was to select natural language terms from the document text as subject keywords, with an option of later replacing those words, stems or phrases by

controlled vocabulary terms. It was Luhn (1957) who first suggested a simple method based on frequency count. Later it was improved by identifying stop-words, stemming, identification of multi-word phrases, and computation of weights (Salton 1989).

With the exponential growth of the World Wide Web, automated subject classification of web pages has become a major research issue. Compared to other types of text documents such as traditional research papers, web pages constitute a special challenge. They tend to be rather heterogeneous: very short with hardly any text, or very long, with titles that are often general ("Home page") or non-existent ("Untitled document"), and metadata that are inconsistent or misused. User-based evaluation of automated classification has been called for but rarely conducted (Ingwersen and Järvelin, 2005).

Organizing web pages into a hierarchical structure for subject browsing has been gaining more recognition as an important tool in information-seeking processes. Usefulness of classification schemes for browsing web resources has been reported but rarely researched by empirical user studies (Vizine-Goetz, 1996; Koch and Zettergren, 1999; Soergel, 2004; Koch *et al.*, 2006). The present paper presents findings from an empirical user study that investigated the performance of an automated classification algorithm on a collection of engineering web pages, in the context of hierarchical browsing. (This has, to the knowledge of the authors, not been conducted before.)

One could distinguish between several different approaches to automated classification (Jain *et al.*, 1999; Moens, 2000; Sebastiani, 2002; Golub, 2006a). In this study a string-matching algorithm is applied, which does not require pre-classified training documents, often unavailable, especially for web pages. The algorithm searches for strings from the Engineering Information (Ei) thesaurus and classification scheme (Milstead, 1995) in text of web pages to be classified. When a string is found, the class which it designates is assigned to the web page. Performance results for the algorithm on a collection of paper abstracts were reported as comparable to state-of-the-art algorithms in machine learning, especially for certain classes (Golub *et al.*, 2007).

It was decided that the Ei classification scheme would be chosen as the target controlled vocabulary. It has been used and maintained in the Compendex database (Engineering Information, 2006). This classification scheme has characteristics that indicate its appropriateness for the task of subject browsing, such as strong design principles and hierarchical structure. In comparison, 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, 48, 76).

Automated classification research using the engineering discipline as a test-bed has not been often conducted. This is due to the fact that the most common approach to automated classification is a machine learning one, where pre-classified document collections are required from which the algorithm "learns". And major document collections used in these experiments neither cover engineering documents, nor do they apply the Ei classification scheme. However, some examples of automated classification utilizing non-mainstream collections exist. The general approach used in this study has been applied earlier in two robot-generated web indexes (DESIRE 2000; Lindholm *et al.*, 2003). The project Bilingual Automatic Parallel Indexing and Classification (Nübel at al. 2002) was aimed at indexing and classifying abstracts from engineering in English and German, using three controlled vocabularies. The INSPEC thesaurus was also applied in several cases (Aitchinson and Harding, 1982; Plaunt and Norgard, 1997). McMahon *et al.* (2004) describe an integrated retrieval system for engineering documents, with one module being a constraint-based classifer. The constraints are mappings between classes and terms and need to be manually built.

The purpose of the present study was to investigate whether it is meaningful to use the Ei classification scheme for browsing, and then, if proven useful, to investigate the performance of the classification algorithm based on the Ei classification scheme and thesaurus.

The paper is structured as follows: in the following section, the classification algorithm and the classification scheme are described (2 Background); in the third section the study design is presented (3 Methodology); in the fourth section (4 Results) the results are analysed and discussed; and, final remarks are given in the last section (5 Conclusions).

2 Background

2.1 Classification algorithm

This section describes the classification algorithm used in the study. The algorithm compares terms from the Ei thesaurus and classification scheme (in further text: Ei controlled vocabulary) to text of web pages to be classified. The Ei controlled vocabulary consists of two parts: a thesaurus of engineering terms, and a hierarchical classification scheme of engineering topics. These two controlled vocabulary types have traditionally each had distinct functions: the thesaurus has been used to describe a document with a number of controlled terms and thus allow as many access points as possible, while the classification scheme has been used to group similar documents together to the purpose of shelving them and allowing systematic browsing.

A major advantage of the Ei controlled vocabulary for automated classification is that thesaurus descriptors are mapped to classes of the classification scheme. These mappings have been made manually (intellectually) and are an integral part of the thesaurus. Compared with captions[1] alone, mapped thesaurus terms provide a rich additional vocabulary for every class. Hence, instead of having only one term per class (there is only one caption per class), there were on average 14 terms per class in the study (see Figure 1 and 2 for an example).

Pre-processing steps of Ei included normalizing upper- and lower-case words. Upper-case words were left in upper case in the term list, assuming that they were acronyms; all other words containing at least one lower-case letter were converted into lower case. The first major step in designing the algorithm was to extract terms from Ei into a term list. It contained class captions, thesaurus terms, classes to which the terms and captions map or denote, and a weight indicating how appropriate the term is for the class to which it maps or which it designates. (Geographical names were excluded on the grounds that they were not engineering-specific.) The term list was formed as an array of triplets:

Weight: **Term** (single word, Boolean term or phrase) = **Class**

Single-word terms were terms consisting of one word. *Boolean terms* were terms consisting of two or more words that must all be present but in any order or in any distance from each other. Boolean terms in this form were not explicitly part of Ei, but were created to our purpose. They were considered to be those terms which in Ei contained the following strings: *and*, *vs.* (short for *versus*), , (comma), *;* (semi-colon, separating different concepts in class captions), *(* and *)* (parentheses, indicating the context of a homonym), *:* (colon, indicating a more specific description of the previous term in a class captions), and -- (double dash, indicating *heading--subheading* relationship). These strings were replaced with *@and* which indicated the Boolean relation in the term. All other terms consisting of two or more words were treated as *phrases*, i.e., strings that need to be present in the document in the exact same order and form as in the term.

The following figure shows two excerpts from the Ei classification scheme and thesaurus. The excerpts found the basis for the creation of term lists (see Figure 2).

Figure 1. Excerpts from the Ei classification scheme and thesaurus

[1] A caption is a class notation expressed in words, e.g., in the Ei classification scheme “Electric and Electronic Instruments” is the caption for class “942.1”.

From the classification scheme:
931.2 Physical Properties of Gases, Liquids and Solids
...
942.1 Electric and Electronic Instruments
...
943.2 Mechanical Variables Measurements

From the thesaurus:
TM Amperometric sensors
UF Sensors--Amperometric measurements
MC 942.1
...
TM Angle measurement
UF Angular measurement
UF Mechanical variables measurement--Angles
BT Spatial variables measurement
RT Micrometers
MC 943.2
...
TM Anisotropy
NT Magnetic anisotropy
MC 931.2

All the different thesaurus terms as well as captions were added to the term list. While choosing all the types of thesaurus terms might lead to precision losses, the decision was to use them all to achieve maximum recall, as shown in a previous paper (Golub, 2006b). In the thesaurus, TM stands for the preferred term, UF ("used for") for an equivalent term, BT for broader term, RT for related term, NT for narrower term. MC represents the main class; sometimes there is also OC, which stands for optional class, valid only in certain cases. Main and optional classes are classes from the Ei classification scheme that have been manually mapped to thesaurus terms and are an integral part of the thesaurus. Based on the above excerpts, the following term list would be created:

Figure 2. Term list created based on excerpts from Figure 1

1: physical properties of gases @and liquids @and solids = 931.2,
1: electric @and electronic instruments = 942.1,
1: mechanical variables measurements = 943.2,
1: amperometric sensors = 942.1,
1: sensors @and amperometric measurements = 942.1,
1: angle measurement = 943.2,
1: angular measurement = 943.2,
1: mechanical variables measurement @and angles = 943.2,
1: spatial variables measurement = 943.2,
1: micrometers = 943.2,
1: anisotropy = 931.2,
1: magnetic anisotropy = 931.2,

The number at the beginning of each triplet is weight estimating the probability that the term of the triplet designates the class; in this example it is set to 1.

The algorithm looks for strings from a given term list in a web page to be classified and if the string (e.g., *magnetic anisotropy* from the above list) is found, the class(es) designating that string in the term list (*931.2* in the example) is(are) assigned to the web page. One class can be designated by many terms, and each time the class is found, the corresponding weight (*1* in the example) is added to a score for the class. The scores for each class are summed up and classes with scores above a certain cut-off (heuristically defined) will be selected as the final ones for the web page being classified.

Findings and setting that had proven best in a previous experiment were applied (Golub *et al.*, 2007). Weights 1, 3, and 4 for single, phrase or Boolean term were multiplied by the weight for the type of class to which the term mapped, 1 or 2 for optional or main class, and by the weight for the type of Ei term (broader 1, narrower 2, preferred 4, related 1, synonyms 3 and captions 4). Also, only those terms based on which correct classes were always derived were included (1,308 terms). Stemming and stop-words removal were not applied. The study also showed that in order to assign a certain class as final, the score of that class had to have

at least 10% of the sum of all the classes' scores (for each web page); in case not a single class with high enough score existed, the one with the highest score was assigned.

As shown in another experiment (Golub and Ardö, 2005), text coming from all the four parts of a web page (title, headings, main text, metadata) should be included in the process of automated classification. The same weights that have previously performed best were applied. The scores of classes found in each part were multiplied with the following weights:

$$\text{Score(class)} = 86 \cdot \text{Score(Title)} + 5 \cdot \text{Score(Headings)} + 6 \cdot \text{Score(Metadata)} + \text{Score(Main Text)}.$$

2.2 Engineering Information classification scheme

The Ei classification scheme is hierarchical and consists of six main classes divided into 38 finer classes which are further subdivided into 182 classes. These are subdivided even further, resulting in some 800 individual classes in a five-level hierarchy. For this study one of the six main classes was selected, together with its subclasses: class 9, *Engineering, General*. The reason for choosing this class was that it covers both natural sciences such as physics and mathematics, and social science fields such as engineering profession and management. The literature of the latter tends to contain more polysemic words than the former, and as such presents a more complex challenge for automated classification. Within the 9 class, there are 99 subclasses; their distribution at the five different hierarchical levels is as follows: 11 classes at the fifth hierarchical level, 67 at the fourth, 16 at the third, and 5 at the second one.

3 Methodology

3.1 Web page collection

The collection was automatically created. For automatically collecting web pages for the study, the Combine focused crawler was used (Ardö, 2007). The term list used by the crawler consisted of the same 1,308 terms used in the classification algorithm (described in section 2). No stemming and no stop words removal were applied. The collection contained 18,895 web pages, crawled in the period between 10 and 15 May 2007. There were 518 seed web pages, taken from the Intute subject gateway, the topic of Engineering General (Intute Consortium, 2006).

Once the pages were crawled, they were classified as described in section 2. On average, 1.5 classes were assigned per web page. No web pages were classified into the top two hierarchical levels, because of the classification principle to assign the most specific class available.

3.2 User study design

The purpose of the study was twofold: to investigate how suitable the Ei classification scheme is for browsing, and how the classification algorithm performs on a harvested collection of web pages. Two major research questions were:

1) Are users able to navigate the Ei classification structure?
2) How well are the web pages classified, as judged by the users?

The first question was investigated by a user study where users browsed the structure when solving four search tasks. For the second question the users were asked to judge the correctness of the automatically assigned classes to documents found in the user study.

The relationship between the two questions was also investigated. This was addressed by conducting a correlation analysis, as well as by analysing respondents' answers from questionnaires.

3.2.1 Experimental setting

The evaluation framework was experimental. In order to get realism in the experiment, four controlled search tasks were developed as background for the browsing and evaluation activities that represented typical information needs that might be encountered in the context of the test collection. The search tasks were developed following the methodology of Borlund (Borlund, 2003). The framework was inspired by previous realistic, interactive user studies (e.g. Nielsen, 2004; Larsen *et al.*, 2006). It comprised the following steps, with the last three ones repeated for each of four search tasks (described in section 3.2.3):

1. Invitation to participate. Participants were recruited through personal contact at the Department of Electrical and Information Technology, Faculty of Engineering, Lund University, advertising the study on mailing lists for courses at the Department and paper adverts on billboards throughout the buildings belonging to the Faculty of Engineering, Lund University. They were allowed to take part if they were undergraduate, graduate, or doctoral students or if had just completed their degree. Each participant was awarded two cinema tickets.
2. Participation consent form. Each participant was asked to sign a participation form which gave information about the study and the participant's role in it.
3. Written instructions: a two-page instructions about the information retrieval system and evaluation criteria.
4. Pre-study questionnaire on participants' background and their previous searching and browsing experience.
5. Search task description.
6. Search session, in which every move was logged. In addition, six participants were asked to "think-aloud" (Lewis and Rieman, 1994, chapter 5), which was video-taped.
7. Post-task questionnaire on participants' certainty of their decisions and general satisfaction.

Each participant was first given the participation consent form and then written instructions on how to conduct the searching. After optional testing of the system and completion of the pre-study questionnaire, the first search task was described and the first searching session began. In the searching session, the participant had to find the class where he/she thought most web pages on the topic of the search task should be. Every click in the browsing tree was logged using home-made software. Once the class was found, the participant evaluated whether the web pages in the class were about the topic of the task. At least top 10 web pages (ranked by descending scores described in 2.) had to be evaluated in a row; at most 40 were offered per screen. The maximum number of web pages evaluated per class was 40, by three participants. Once the participant decided to be finished with the task, he/she filled-in a post-task questionnaire.

The language of the study was English. The vast majority of web pages were also in English. The study took place in the period between 21 May and 5 June 2007. It was conducted at one of the Department's computer rooms, one participant per computer. Six of the participants were video-taped in the researcher's office. The researcher was always present and available for help or clarifications. Each session was predicted to last one hour, as stated in the invitation to participate and the participation consent form.

3.2.2 User interface

Since, to the authors' knowledge, an operative information system employing the full-scale Ei classification scheme as a browsing structure did not exist, a simple home-made interface was created (Figure 3). It consisted of two parts: in the upper part of the screen, a clickable hierarchical browsing tree of the Ei classes was provided, and in the lower part web pages classified in the class clicked on were listed, in a descending relevance order based on classification scores. In the upper part of the screen also most important instructions were given; detailed instructions were provided on a separate sheet of paper. For each web page

there was an automatically extracted title, automatically extracted sentences ("Summary"), hyperlink to the original web page and a small evaluation form. In the evaluation form the participant was asked to judge whether the web page is about/concerns/deals with the topic of a given task. Four options were available: "correct", "partly correct", "incorrect" and "impossible for me to say".

The retrieval system and the questionnaires were pilot-tested by two additional participants. Based on their input, several minor changes on the user interface were implemented.

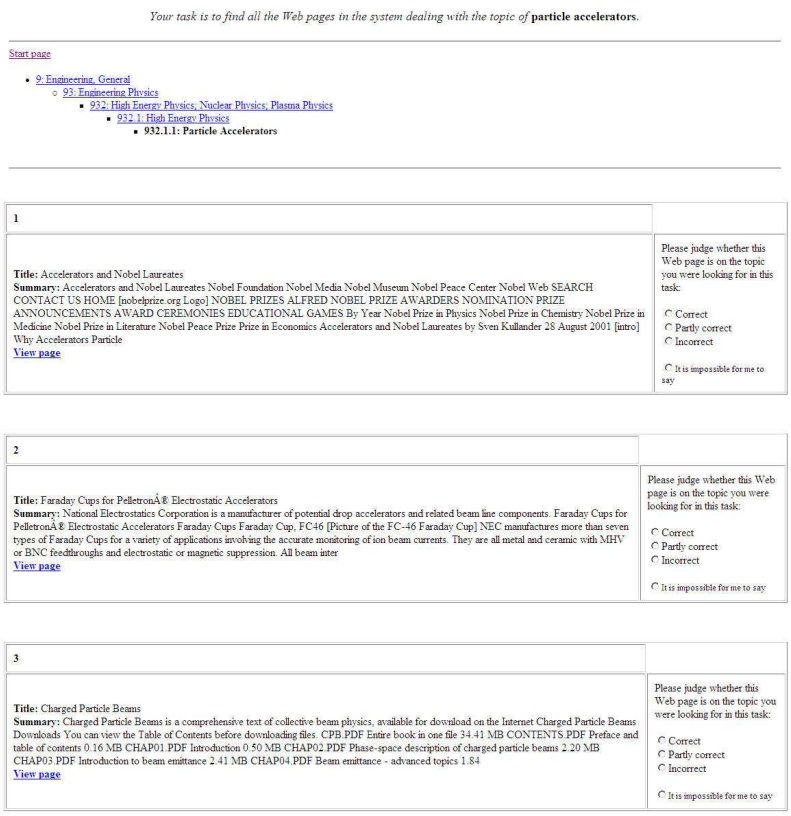

Figure 3. Screen shots of main parts of the user interface: the task definition, browsing tree and top-ranked web pages with evaluation forms.

3.2.3 Search tasks

Each participant was given four search tasks. A similar number of tasks was used in related studies (e.g., Nielsen, 2004; Larsen *et al.*, 2006). The tasks were presented in one of 24 rotated sequences, as recommended by Borlund (2003).

In each task the participant was to find web pages on an assigned topic (cf. Ingwersen and Järvelin, 2005, 73 for tasks in information retrieval studies). Exact formulations of the four tasks were as follows:

1) "*Your task is to find all the web pages in the system dealing with the topic of* **particle accelerators**."

2) "*Your task is to find all the web pages in the system dealing with the topic of* **magnetic instruments**."

3) "*Your task is to find all the web pages in the system dealing with the topic of* **differentiation and integration."**

4) *"Your task is to find all the web pages in the system dealing with the topic of* **professional organizations in the field of engineering**."

Two tasks were in the basic sciences as applied in engineering, and two in general engineering; two were at fifth, and two at the fourth hierarchical level; in two topic names were the same as class captions and two were entirely different:

- **Task 1** was in the field of physics, on the topic of *particle accelerators*. The class was at the fifth hierarchical level (932.1.1), the class caption was the same as the topic name, and words from the topic name did not exist anywhere in captions of higher level classes.
- **Task 2** was in the field of instruments and measurements, on the topic of *magnetic instruments*. The class was at the fourth hierarchical level (942.3), the class caption was the same as the topic name. One word (*instruments*) from the topic name was part of captions at second (94 *Instruments and measurement*) and third (942 *Electric and Electronic Measuring Instruments*) hierarchical levels.
- **Task 3** was in the field of mathematics, on the topic of *differentiation and integration*. The class was at the fourth hierarchical level (921.2), the class caption (*Calculus*) was entirely different from the topic name, and words from the topic name did not exist anywhere in the higher level class captions.
- **Task 4** was in the field of engineering profession, on the topic of *professional organizations in the field of engineering*. The class was at the fifth hierarchical level (901.1.1), the class caption (*Societies and Institutions*) was entirely different from the topic name, and one word (*professional*) from the topic name existed in its noun form at second (901 *Engineering Profession*) and third hierarchical levels (901.1 *Engineering Professional Aspects*).

3.2.4 Data collection and analysis methods

In order to investigate browsing (research question 1), quantitative data were collected by logging the browsing steps and determining if correct classes were found. Participants' browsing steps and selected classes were compared against a standard reference ("ideal") browsing path and a standard reference class for each search topic. These were pre-determined by the researcher, by simply looking at the whole browsing tree and identifying what the best matching class for each search topic would be, and what the shortest route to the class would be. For each task, the shortest possible (standard reference) browsing path and class were known and the participants' steps were compared against them.

For the classification part of the study (research question 2), quantitative data were collected through user assessment of the correctness of the assigned classification code. For each web page listed under a class selected as the most appropriate for the given task, a form was offered with four options from which to choose: "correct", "partly correct", "incorrect" and "impossible for me to say". According to the written instructions, these were to be chosen in the following cases:

- Correct – if the web page is about the topic;
- Partly correct – if the web page can be considered to be on the topic, but is mixed with other topics; and,
- Incorrect – if the web page has absolutely no relation to the topic.

The participants were asked to avoid the option "Impossible for me to say" and only use it in cases when the web document content was not available, e.g., if a Web server was down. Apart from indicating different level of topic coverage (cf. weighted indexing in Lancaster, 2003, 187-188), the three options are also analogous to related experiments were relevance of documents is assessed by use of three relevance levels, "relevant", "partly relevant" and "irrelevant" (e.g., Nielsen, 2004).

Furthermore, evaluations of both standard reference classes for each task and others selected by users as *the* class for a task were examined.

For both parts of the study, a post-task questionnaire on participants' certainty of their decisions and general satisfaction was to be filled-in after each task. In addition, clarifying, qualitative data were collected for six participants by observation based on the "think-aloud" protocol. After completing their search tasks, all the participants were asked if they had any comments, which were also recorded and analysed. At this stage, comments were received from 12 participants. Through post-task questionnaires, 31 comments were collected, submitted by 16 participants.

Based on post-task questionnaires, relation between browsing and classification correctness was investigated. For each question, 159 answers were collected. Correlations were calculated between every pair of questions that could indicate an influence of classification correctness on browsing and vice versa. As the answers were of ordinal variable type, Spearman's rank correlation was used (Vaughan, 2001, 140-143). The calculations were conducted in Matlab, where the original data were given as input to a formula for direct calculation of Spearman's *rho* values.

Information on participants' background and their previous searching and browsing experience were collected in a pre-study questionnaire.

3.2.5 Participants

There were 40 participants, students and researchers in the field of engineering. The participants were selected randomly: they were the first 40 people who agreed to take part in the study.

Information on participants' background and their previous searching and browsing experience was collected through the pre-study questionnaire.

- The majority (85%) had very good or excellent knowledge of English.
- All participants had at least four years of online searching experience.
- The majority (87.5%) were between 20 and 30 years old.
- The majority (88%) were male.
- The majority (86%) were taking or have completed their Master's degree in the field of computer engineering.
- The majority (90%) claimed they generally found what they were looking for on the World Wide Web.
- On average they used search engines once or twice a week (3.8 on a scale from 1 to 5 where 1 stands for "Never", 2 for "Once of twice a year", 3 for "Once or twice a month", 4 for "Once or twice a week", and 5 for "One or more times a day").
- On average they used professional information services such as library catalogues and Lund University's service providing free access to commercial databases once or twice a year (1.8 the former, 1.6 the latter), and engineering-specific database Compendex (also freely available for Lund University's students and researchers), hardly ever (1.1). This is in significant contrast to the use of search engines.
- On average they used hierarchical directory-style browsing of, e.g., search engines or other information databases once or twice a month (2.5).

The group of users was selected from the population of engineering students or young engineers who had recently acquired their degrees. All people who fulfilled this criterion and expressed their interest in taking part in the study were accepted. While these were primarily users whose first language was Swedish, they had very good or excellent knowledge of English, which implies that the results of the study can be in large part generalized. However, the language might have influenced several participants in their decisions, such as the one suggested in Section 4.1.1 (Task 3).

4 Results

4.1 Browsing

The suitability of the Ei classification scheme for browsing was evaluated by two measures: the number of participants finding the standard reference class predefined for each of the four tasks; and, the number of individual classes visited before reaching the final class. Comments received from the participants were also analysed. The influence of automated classification correctness on browsing decisions is discussed in section 4.3.

4.1.1 Analysis based on browsing steps

As described earlier (section 2.4.1), the participants were instructed to find the class they consider most appropriate for the task at hand, and evaluate whether the web pages listed under the class concerned the topic of the task. In several cases, the participants choose to evaluate web pages from more than one class.

On average, the majority (29 out of 40 participants) found the right class. Approximately, two other classes per task were considered correct by at least two participants.

In Table 1 responses from the post-task questionnaire related to browsing are presented. The results are reported in separate columns for those who found the right class ("right class found") and for those who did not ("right class not found"). On a scale from 1 to 3, where 1 stands for "not at all", 2 for "somewhat" and 3 for "very", participants who found the right class reported on average for all the four tasks that it was easy (2.3) to find the right class ("easycat") and that they were rather certain they found it ("certaincat") (2.6). Those who did not find the right class were less sure they found it (1.7) and for them it was less easy to find an appropriate class (1.9). Both groups reported that they were somewhat familiar with the topics (2.1).

Table 1. Results from post-task questionnaires related to browsing.
The scale is from 1 to 3, where 1 stands for "not at all", 2 for "somewhat" and 3 for "very".

	right class found			right class not found		
task	easycat	certaincat	familiar	easycat	certaincat	familiar
1	2.2	2.9	1.8	1.8	2.0	1.9
2	2.7	3.0	1.9	2.0	1.6	2.2
3	2.4	2.5	2.7	1.7	1.4	2.2
4	1.9	2.0	2.0	1.9	1.7	1.9
average	2.3	2.6	2.1	1.9	1.7	2.1

For task 2 and 3, the standard reference browsing path takes four steps and for task 1 and 4 the path takes five steps. On average, the participants who found the right class took 15 steps; all participants made on average 16 steps. Browsing in each task is discussed separately below.

Task 1: *particle accelerators* (932.1.1)

In Task 1 the shortest possible number of steps was five (including the step in which one decided that he/she reached the right class). There were six participants who followed this shortest path. The majority (21) took up to 15 steps to come to the standard reference class. On average participants who found the standard reference class took 16 steps; all participants took 19 steps. An example of a 15-step sequence taken by one participant is given below:

93 → 932 → 932.2 → 932 → 93 → 931 → 931.3 → 93 → 9 → 94 → 9 → 93 → 932 → 932.1 → 932.1.1

As seen from Table 2, the first browsing step the 70% majority took was correct – 93 *Engineering Physics*. Eight participants chose 94 *Instruments and Measurement* and four 90 *Engineering, General*. Of those who took the correct first step, the majority chose the correct second step (82.1%). The weakest point was the third standard reference step, choosing a

specific class within the broad area of 932 *High Energy Physics; Nuclear Physics; Plasma Physics* – only 43.5% made the right decision. Most of them were not sure if *particle accelerators* belonged to 932.1 *High Energy Physics*, 932.2 *Nuclear Physics* or 932.3 *Plasma Physics*. Since the three classes at the fourth hierarchical level seem quite clear, each representing one of the three concepts from their broader class 932, this could be attributed to the fact that the participants were less than somewhat familiar with the topic of the task (1.8 and 1.9, see Table 1).

Table 2. Standard reference browsing steps for Task 1. Percentage in "step taken by" is calculated in relation to the number of participants taking the preceding standard reference step.

	ideal sequence	step taken by
step 1	93 Engineering Physics	70.0%
step 2	932 High Energy Physics; Nuclear Physics; Plasma Physics	82.1%
step 3	932.1 High Energy Physics	43.5%
step 4	932.1.1 Particle Accelerators	80.0%
step 5	confirmed	87.5%

There were 31 participants who found the right class. Six other classes were deemed correct by at least one participant. Classes chosen by at least two participants were the following:

- 932 *High Energy Physics; Nuclear Physics; Plasma Physics*, which is a correct but not the most specific class in the hierarchy. Considering a broader class as correct, especially when relevant resources were discovered, is a defendable error;
- 931.3 *Atomic and Nuclear Physics*, which is wrong. This class's broader class 931 *Applied Physics Generally* is also wrong, but the first step class (93 *Applied Physics Generally*) is correct;
- 932.2.1 *Fission and Fusion Reactions*, which is wrong. This class's broader class 932.2 *Nuclear Physics* is also wrong, but the first and second step classes are correct.

Thus, all the participants chose the correct second hierarchical level, 93 *Engineering Physics*. The reason why they chose different classes within physics could be attributed to the fact that they were less familiar with the topic. The participants who found the right class were very certain that they found it (2.9). This could be partly explained by the fact that the class caption was the same as the topic name. Participants who did not find the right class were only somewhat sure (2.0); for them also finding the class was less easy (1.8 on the scale) than for those who found it (2.2). Both groups were a bit less than somewhat familiar with the topic from before (1.9 for those who did not find it and 1.8 for those who did).

Task 2: *magnetic instruments* (942.3)

In Task 2 the shortest possible number of steps was four (including the step in which one decided that he/she reached the right class). There were 18 participants who followed this shortest path. The majority (24) took up to six steps to find the standard reference class; on average eight steps were taken. An example of a six-step sequence taken by one participant is given below:

94 → 943 → 943.3 → 94 → 942 → 942.3

As seen from Table 3, more than half of the participants (60%) took the correct first step; 13 participants chose 93 *Engineering Physics* and 3 chose 90 *Engineering, General*. Of those who took the correct first step, the majority chose the correct second step; those who did not, chose 943 *Mechanical and Miscellaneous Measuring Instruments* which could be, because of

the word *miscellaneous*, considered justifiable. The majority chose the correct third step, and also confirmed the class to be final.

Table 3. Standard reference browsing steps for Task 2.

	ideal sequence	step taken by
step 1	94 Instruments and Measurement	60.0%
step 2	942 Electric and Electronic Measuring Instruments	79.2%
step 3	942.3 Magnetic Instruments	94.7%
step 4	confirmed	94.4%

There were 35 participants who found the right class. Eight other classes were deemed correct by at least one participant. The class chosen by at least two was 931.1 *Mechanics*. This class is incorrect, although its higher levels 931 *Applied Physics Generally* and 93 *Engineering Physics* could be considered correct to some degree. The participants who found the right class were very certain that they found it (3.0 on the scale, as seen from Table 1). This could be partly explained by the fact that the class caption was the same as the topic name. Participants who did not find the right class were far less sure (1.6). For them also finding the class was less easy (2.0) than for those who found it (2.7). Both groups were somewhat familiar with the topic from before (2.2 for those who did not find it, and 1.9 for those who did).

Task 3: *differentiation and integration* (921.2)

In Task 3 the shortest possible number of steps was four (including the step in which one decided that he/she reached the right class). Eight participants followed this shortest path. The majority (22) took up to 14 steps to come to the standard reference class; on average, 15 steps were taken by those was found the standard reference class, and 18 by all. An example of a 14-step sequence taken by one participant is given below:

92 → 921 → 921.2 → 921 → 921.3 → 921 → 921.5 → 921 → 92 → 922 → 922.2 → 92 → 921 → 921.2

Table 4. Standard reference browsing steps for Task 3.

	ideal sequence	step taken by
step 1	92 Engineering Mathematics	85.0%
step 2	921 Applied Mathematics	91.2%
step 3	921.1 Calculus	51.6%
step 4	confirmed	37.5%

As seen from Table 4, the first browsing step the majority took was correct (85%) – 92 *Engineering Mathematics*. Of the remaining six, two went to 90 *Engineering General*, two to 91 *Engineering Management* and two to 94 *Instruments and measurement*. Of those who took the correct first step, the majority chose the correct second step (91.2%). A weak point was the third step, choosing a specific class within the broad area of 921 *Applied Mathematics*: only a tight majority picked the right class (51.6%). The weakest point was coming to the right class – only 37.5% decided it was the standard reference class, while others went one level up, to class 921 *Applied Mathematics*. The authors believe that the reason for the latter two could be attributed to the fact that the class caption was entirely different from the topic name. Also, the participants not finding the standard reference class were not very familiar with the topic (2.2 on the scale, see Table 1). Another reason could be that the participants were mainly Swedish and did not take mathematics courses in English, and the English word *calculus* shares little more than etymology with the Swedish word *kalkyl* which in common usage means *calculation* (the Swedish term for *calculus* is *analys*).

There were 28 participants who found the right class. Nine other classes were deemed correct by at least one participant. Classes chosen by at least two of them were the following:

- 921.6 *Numerical Methods* which is wrong, but its broader class 921 *Applied Mathematics* is correct; and
- 921.4 *Combinatorial Mathematics* which is wrong, but also has its broader class 921 *Applied Mathematics* correct.

Thus, all participants have chosen correct second and third hierarchical levels (92 *Engineering Mathematics* and 921 *Applied Mathematics*). The authors believe that the reason why some of them did not chose the standard reference class at the fourth hierarchical level could be the same as above for the fourth browsing step. The participants who found the right class were between somewhat and very certain they found the right class (2.5 on the scale, as seen from Table 1). The certainty level is high, although a bit lower than in the first two tasks. This may be explained by the fact that the class caption was different from the topic name. Participants who did not find the right class were rather unsure (1.4 on the scale). For them finding the class was also less easy (1.7) than for those who found it (2.4). The group who found the standard reference class was quite familiar with the topic from before (2.7), while the group who did not find the standard reference class was somewhat familiar (2.2).

Task 4: *professional organizations in the field of engineering* (901.1.1)

In Task 4 the shortest possible number of steps was five (including the step in which one decided that he/she reached the right class). There were five participants who followed this shortest path. On average 19 steps were taken by those who found the standard reference class, and also by all. An example of a 19-step sequence taken by one participant is given below:

9 → 91 → 912 → 912.2 → 9 → 90 → 901 → 901.1 → 901.1.1 → 901.1 → 90 → 901 → 9 → 91 → 913 → 9 → 90 → 901 → 901.1 → 901.1.1

This example also demonstrates how the participant found the right class but continued looking at other classes until he/she made the decision he/she was most sure of. Reasons why this was the case need further investigation.

As seen from Table 5, the first standard reference browsing step was taken by half of the participants; the second half chose class 91 *Engineering Management*. This may be explained by the nature of the topic, considered to be partly correct. All those who took the standard reference first step, chose the correct second step. While the majority also chose the correct third and fourth step, only half of those who came to the right final class realized it was the right one. The reason for this could be that the class caption was entirely different from the topic name.

Table 5. Standard reference browsing steps for Task 4.

	ideal sequence	step taken by
step 1	90 Engineering, General	50.0%
step 2	901 Engineering Profession	100.0%
step 3	901.1 Engineering Professional Aspects	70.0%
step 4	901.1.1 Societies and Institutions	71.4%
step 5	confirmed	50.0%

There were 20 participants who found the right class. Nine other classes were deemed correct by at least one participant. Classes chosen by at least two participants were the following:

- 912.2 *Management*, which could be considered correct, although there is a class that describes the topic better;
- 901.1 *Engineering Professional Aspects*, which is correct but not the most specific class that can be found in the classification scheme. Considering a broader class as correct, especially when relevant resources were discovered, is a defendable error;

- 901.3 *Engineering Research*, which could be considered correct, although there is a class that better describes the content; and
- 912.1 *Industrial Engineering*, which is also somewhat related to the topic of the task.

Reasons for choosing these different classes could be attributed to the fact that the topic of this task can belong to more than one strict class. Participants who found the right class were somewhat sure they found the standard reference class (2.0 on the scale), less than for the other three tasks. This could be explained by the nature of the topic, the fact that the class caption was entirely different from the topic name, and also by the fact that the top ranked web pages in this class were evaluated to be between partly correct and incorrect, worse than in any other task. Comments by participants confirmed that this task was more ambiguous than others, e.g., "This one seems like a broad topic, it could include anything that works with engineering professionally, quite a lot" or "All companies are also at least semi-professional so I feel I can pretty much go to any category and still find things."

Participants who did not find the right class were less sure they reached it (1.7). Both groups reported that finding the right class was somewhat easy (1.9), but on average it was more difficult than in any other task. The familiarity with the topic was reported to be medium by both groups (2.0 where correct and 1.9 where incorrect).

4.1.2 Analysis based on comments

Since comments were not obligatory, this section only provides indications. Several participants said that they preferred searching to browsing, and some of them provided reasons such as the following ones:

- "I prefer searching because here you always need to go up and down";
- "I think you need to know something about the topic before you start using the tree; the hierarchy helps if you know a little bit, but if you have no clue…"

These issues could be dealt with by enhancing the hierarchical interface by, for example, adding a search box for words in class captions with synonym search (easily provided for Ei since the classes are mapped to thesaurus terms), and returning the hierarchical tree expanded around the class in which caption the search term is found. If a term searched for is contained in, say, two different contexts, returning the two hierarchical trees would serve as a disambiguation device and help the user chose the exact meaning he/she is looking for. The suggestion to allow for searching for words from class captions was provided by quite a few participants as well. Another suggestion they made was to provide some kind of a support for explaining classes, e.g., describing each caption with what it contains or adding an expand box with what is below the class.

Several participants expressed that they liked browsing. One said that he/she had never tried browsing before. By the time he/she arrived to the last task, he/she was happy with the experience and expressed that it had been pleasing. Another said he/she didn't really like the hierarchy, but believed it could be useful once used to it.

One participant complained that for two classes it was not obvious that they would be where they were; however, the fact that he/she did find them shows that it is possible to find one's way through the Ei structure although every individual would probably structure subjects differently. Two participants commented that class 94 *Instruments and Measurement* represents an application area, while others are scientific, and that there are overlaps in subjects between them – e.g., *instruments* could be part of physics as well. Overlaps in topic representations in the classification scheme were reported as an issue by others as well, but could be dealt with a search entry into a synonym list of class captions (already part of the Ei thesaurus). Moreover, overlaps in topics exist in disciplines themselves and a good classification scheme should reflect such overlaps. Another suggestion was to exclude the words *general* and *engineering* from captions of second-level classes as he/she perceived

them as redundant. Several wanted to see more specific subclasses; the reason could be that there were too many web pages in one class, more than 40 in most of them.

4.2 Automatically assigned classes

In the second part of the study, the correctness of automatically assigned classes was analysed. The most common approach is by comparing automatically assigned classes against human-assigned ones. For web pages there are few collections with human-assigned classes. To the authors' knowledge, there is one maintained collection using Ei classes, Intute subject gateway on engineering (Intute Consortium, 2006). However, web pages in this collection are classified mainly into top hierarchical levels, or six classes altogether: 900, 910, 920, 930, 940, 901.2. Since the aim was to study how the algorithm performs also at third, fourth and fifth hierarchical levels, this collection did not suffice. Moreover, the problem of documents' "aboutness" has been much discussed in the literature and the need for evaluating automated classification by end users has been proposed but seldom conducted (cf. Ingwersen and Järvelin, 2005).

4.2.1 Automatically assigned classes against human-assigned ones

Since web pages were automatically crawled and classified, no pre-existing human-assigned classes were available. Of the 518 web pages that were classified both by the Intute subject gateway (Intute Consortium, 2006) and by the algorithm, 320 of them (62%) were put in the same class as in the Intute subject gateway. Because of the small sample and because Intute has most web pages only at the top two hierarchical levels, further comparison was not conducted.

4.2.2 Automatically assigned classes as judged by the user study participants

Table 6 shows the number of web pages evaluated in different tasks. In standard reference classes there were on average 40 different web pages evaluated per task with 10 different participants evaluating each web page. There were in total 36 evaluations or 23 different web pages that were deemed as "impossible for me to say" by at least one participant. Four web pages were deemed as "impossible for me to say" by two or more participants. One of them was "under construction", two had very little text, and one was extremely long (178 pages if printed). Some of them were also not in English. Because of the small number of "impossible to say" decisions, they were not counted in tables following this one.

Table 6. Number of evaluations and evaluated web pages.

	task 1	task 2	task 3	task 4	in total
Number of evaluations	620	644	593	617	2474
Number of different Web pages evaluated	170	154	172	176	672
Number of evaluations in the ideal class	392	497	417	297	1603
Number of different web pages in the ideal class evaluated	40	31	40	40	151

Table 7 shows for each task the standard reference class and averaged evaluations for 1) all the evaluated web pages, and 2) top 10 ranked web pages that were at the same time most frequently evaluated. The scale used in this part of the study was from 1 to 3, where 1 stands for "correct", 2 for "partly correct" and 3 for "incorrect". The differences between all and top 10 evaluated web pages do not seem to be significant, but are a little better for the top 10 pages, which is in accord with the fact that the higher ranked ones should be more correct. The fact that the differences are not very significant implies that the algorithm performs equally well for resources listed further down on the page. Neither were there significant differences when comparing evaluations made by participants who were very certain that their judgements were accurate ("only when certain") against the average of all participants ("different certainty").

Table 7. Correctness of automatically assigned standard reference classes.

		different certainty		only when certain		
task	correct class	for all evaluated web pages	for top 10	for all evaluated web pages	for top 10	average
1	932.1.1	1.9	1.8	1.8	1.8	1.8
2	942.3	2.0	1.8	2.1	1.9	2.0
3	921.2	2.0	2.0	2.0	2.0	2.0
4	901.1.1	2.6	2.5	2.6	2.5	2.6
average	n/a	2.1	2.0	2.1	2.1	2.1

As with browsing, the performance of the algorithm as judged by the participants differs between the tasks. Based on 1,603 evaluations of 151 different web pages in standard reference classes (last two rows in Table 6), the top ranked web pages in each of the four classes were on average deemed partly correct (2.1 on the scale). Best classification was achieved for Task 1 (1.8), and the worst one for Task 4 (2.6). This was confirmed by comments given by several participants who said that web pages in Task 1 seemed quite good but that they were confused with web pages in Task 4.

Table 8. Correctness of automatically assigned other classes.

task	other classes	number of participants choosing the class	number of evaluated web pages	for all evaluated Web pages (different certainty degrees)
1	932	5	104	2.3
	931.3	5	54	2.6
	932.2.1	3	30	2.7
2	931.1	3	44	2.2
3	921.6	4	50	2.8
	921.4	2	23	2.4
4	912.2	8	119	**2.3**
	901.1	7	108	**1.9**
	901.3	4	48	**2.1**
	912.1	2	11	2.2

While the majority chose the standard reference class, it was also important to see how the participants who chose other classes as correct, evaluated web pages put in those classes. Table 8 reports evaluations of web pages for non-standard reference classes chosen by at least two participants.

Table 9. Results from post-task questionnaires related to correctness of classes and general experience.

	right class found			right class not found		
task	easyclass	certainclass	experience	easyclass	certainclass	experience
1	2.3	2.4	2.4	1.9	2.3	1.7
2	2.3	2.4	2.4	2.0	2.0	1.6
3	2.3	2.4	2.3	1.6	1.7	1.4
4	2.4	2.3	2.0	2.0	2.0	1.8
average	2.3	2.4	2.3	1.9	2.0	1.6

In Tasks 1, 2 and 3 topics of web pages in non-standard reference classes were deemed more wrong than in the standard reference class. In Task 4, however, evaluations of web pages in the non-standard reference classes were deemed more correct than in the standard reference class, the best one being its broader class, 901.1. This could be partly explained by

the fact that this topic could be considered to fit in several classes (cf. Task 4 in 3.1.1), and as such is harder to automatically classify and judge.

Table 9 shows results from the post-task questionnaires related to correctness of classes as well as general experience ("experience"). The results are reported separately for those who found the right class ("right class found") and for those who did not ("right class not found"). On a scale from 1 to 3, where 1 stands for "not at all", 2 for "somewhat" and 3 for "very", participants who found the right class reported on average for all the four tasks that it was easy (2.3) to decide whether the web pages in the selected class were on the topic of the task ("easyclass") and that they were certain of their evaluation indicating whether web pages in the class were on the topic of the task ("certainclass") (2.4). On a scale from 1 to 3 where 1 stands for "frustrating", 2 for "neutral" and 3 for "pleasing", this group deemed the whole experience to be between neutral and pleasing (2.3). Those who did not find the right said that it was less than somewhat easy to decide whether web pages in the selected class were on the topic of the task (1.9) and were less certain of their evaluation indicating whether web pages in the class were on the topic of the task (2.0). They considered the whole experience to be somewhere between 'frustrating' and 'neutral' (1.6).

The fact that evaluations between the four tasks differed is in line with a previous study, where the algorithm's performance was tested on a pre-classified collection of research abstracts (Golub *et al.*, 2007). The study showed that certain classes performed better than others. Table 10 presents precision and recall for the four classes, but measured on the collection of paper abstracts. While precision is almost total for all the four classes, recall is weakest for class 901.1.1 (Task 4), which can be attributed to the fact that only one term exists for this class on the term list. Also, most terms designating the other three classes are rather field-specific and thus less ambiguous than the term designating class 901.1.1.

Table 10. Performance for the four classes on the collection of paper abstracts (described in Golub *et al.*, 2007), with the same parameters as in this study.

class	term	instances found	precision	recall
932.1.1	accelerators @and betatron	1	0.97	0.08
	electron sources	1		
	particle beams	3		
	accelerators @and electrostatic	1		
	accelerators @and magnets	4		
	accelerator magnets	1		
	accelerators @and targets	4		
	electron accelerators	2		
	accelerators @and synchrotron	3		
	synchrotron x-ray radiation	2		
	storage rings	11		
	particle beam dynamics	2		
	synchrotron ultraviolet radiation	1		
	linear accelerators	2		
942.3	gradiometers @and magnetic	7	1	0.12
	compasses	1		
	magnetometers	35		
	fluxmeters	2		
921.2	kinetic theory of gases	1	1	0.03
	differential relational calculus	1		
	differentiation @and calculus	1		
	integral equations	12		
	maxwell's equations	9		
901.1.1	societies @and institutions	1	1	0.001

4.2.1 The problem of "aboutness"

The challenge of identifying the aboutness of documents has been much discussed in the literature. This is related to the quality of indexing. According to Lancaster (2003, 85), an indexing "failure" could occur in the conceptual analysis phase of indexing, where a topic of user interest is not recognized or is misinterpreted, and in the translation phase, where not the most specific term gets used or the term chosen is inappropriate. In this study there were cases when one web page was at the same time evaluated as correct, incorrect and partly correct. For top 10 pages in all the 4 tasks, 26 pages were at least once evaluated as correct, incorrect and partly correct; 11 pages were evaluated with two different values; and, only 3 pages were evaluated with the same value. Based on "think-aloud" sessions and participants' comments, reasons behind their decisions and such big discrepancies between evaluations were analysed:

- There were three cases implying that some participants used only summaries, in spite of the clearly provided instruction "In order to evaluate whether the web page is about the topic given in the task, please open and look at the page ('View page') because it would be incorrect to judge only based on the Title and Summary". An example of a comment implying they used only summaries was "…this task was a little bit hard to extract information from web sites because the researching subject wasn't described in web sites' summaries…"
- Most differences between how people judged one and the same web page occurred due to mixing web page's topic with its genre. There are many web pages offering not just factual information on a topic, but also (or instead) describe a related software, provide a search engine, or sell products like magnetic instruments. While according to instructions, "incorrect" was supposed to be chosen only when the web page had absolutely no relation to the topic, some judged commercial web pages as incorrect. E.g., a participant leaving a comment "a lot of commercial websites and almost no didactic website" on average evaluated web pages in Task 3 as 2.4 (between partly correct and incorrect), while the average for that task was 2.0 (partly correct).
- Task interpretation. Some evaluated a web page based on whether it had anything to do with the topic at hand, as instructed, while others based it on their own task interpretation and introduced criteria such as usefulness and quality. E.g., a participant saying "this page could be useful if we know something but for a beginner, no" judged that web page as incorrect. Or, a web page providing only a definition on compasses was judged by one participant as partly correct because "it provides only superficial information".

These findings illustrate the problem of aboutness in general, and should be considered when analysing the findings concerning classification correctness.

4.3 Browsing and classification correctness

Finding one's way through the browsing tree tends to be related to whether web pages are classified in appropriate classes. This is supported by participants' comments:

- "I am not sure I found the right category for professional organizations". Actually this participant did find the right class, but was unsure of that because web pages in the class were by that participant evaluated as mostly incorrect (2.6 on average). This topic belongs to Task 4 where web pages were by all deemed to be least correct.
- "Most of them are incorrect so I think I chose the wrong category". This participant really did not find the right class, so in this case web pages were correctly indicating that he/she should look for another class.
- Another participant said: "It is easier with search bar, this is quite frustrating; lot's of useless web pages". In this case the right class was not found, but the comment indicates

how the correctness of automated classification influenced the participant's preference for searching.

Based on each participant's post-task questionnaire answers for every task (159 per question) the following significant correlations were recognized with probability above 95% (Sheskin, 2000, Table A18):

- Certainty that the right class was found and certainty of one's evaluation whether web pages in the found class were on the topic (Spearman correlation coefficient is 0.33).
- Easiness to find the right class and easiness to decide whether web pages in the found class were on the topic: the correlation is (Spearman correlation coefficient is 0.31).
- Certainty that the right class was found and easiness to decide whether web pages in the found class were on the topic (Spearman correlation coefficient is 0.35).

This indicates that the success of browsing is related to the degree to which web pages have been correctly classified.

5 Conclusions

The study was to investigate performance of an automated classification algorithm on a collection of engineering web pages, in the context of hierarchical browsing. Two major research questions were whether users were able to navigate the Ei classification structure and how correctly were web pages in a certain class classified. The study involved 4 tasks and 40 participants. The participants had a very good knowledge of English, at least four years of online searching experience, frequent usage of search engines, once or twice a month they used hierarchical browsing and were generally finding the desired information.

The study showed that the Ei classification scheme is generally well suited for browsing. The majority of participants found the right class, they reported that it was quite easy finding it and were quite certain they found the right class. Also, those who found standard reference classes deemed the whole user study experience between neutral and pleasing. This was the case in spite of the fact that the participants on average made 15 steps to reach the standard reference class, while the shortest browsing path would take only 5. However, the number of browsing steps needs to be put in relation to at least two other factors: examples have shown that some participants have systematically looked at a number of other classes to make sure they reached the most appropriate one; and, the hierarchical tree showed only the path to the last class clicked on, preventing distant jumps. Other possible factors could be inadequacy of the classification scheme and participants' unfamiliarity with it. Inadequacy of the classification scheme was indicated as to the following: captions contain redundant words like *engineering*; division of subject areas is not very logical: basic sciences, mathematics and physics, as applied in engineering, are at the same hierarchical level as *Instruments and Measurement* and *Engineering, General*; and, class *Engineering, General* contains a mixture of topics such as engineering profession, graphics and libraries. Exact reasons behind taking longer paths than required need to be further studied.

Majority of the participants selected correct second and third hierarchical level classes in all the four tasks. More wrong classes were chosen at the fourth hierarchal level, which could be explained by: 1) participants' unfamiliarity with the subject at the required specific level; and, 2) class captions being different from topic names. The latter is confirmed by the fact that approving the right class arrived at as correct was problematic in the two tasks in which class captions were entirely different from topic names. For those two tasks also lower certainty levels were reported. The lowest certainty level was obtained for one of those two tasks in which web pages were judged as more incorrect than in other tasks, another possible contributing factor. Also, the classification scheme could have for some reason been inappropriate. Exact reasons need to be further investigated.

Top ranked web pages in each of the four classes were on average deemed partly correct. A major problem with determining whether a web page is in the right class or not is that there were large differences among participants in their judgements – a number of web pages were evaluated as correct, partly correct and incorrect by different participants. A major reason is probably the reported problem of aboutness" and related subjectivity in deciding which topic a document is dealing with. Other factors were also recognized in the study:

- Some participants, despite the instructions, based their evaluations only on summaries, instead of full-text web pages;
- Others interpreted tasks more narrowly and evaluated web pages based also on other criteria such as quality and usefulness;
- It was hard to evaluate web pages' topicality when there was hardly any or very much text.

As with browsing, evaluations between the four tasks differed. This is in compliance with previous results of the algorithm's performance, based on a pre-classified collection of research abstracts, where it was shown that certain classes have better performance than others (Golub *et al*, 2007).

Although the classification of web pages was on average judged as only partly correct, and while there is evidence that correct placement of web pages and browsing success are related, the majority of the participants were able to navigate the Ei classification structure well and deemed the whole experience to be between neutral and pleasing.

Several improvements for browsing have been identified:

- Describing class captions and/or listing their subclasses from start;
- Allowing for searching for words from class captions with synonym search (easily provided for Ei since the classes are mapped to thesauri terms);
- When searching for class captions, returning the hierarchical tree expanded around the class in which caption the search term is found; and
- Because automatically-produced summaries could be misleading, avoid presenting them until their quality is sufficiently improved.

The need for several improvements of classification schemes was indicated:

- Follow consistent division principles when building classification structures;
- Modify captions so that they better reflect concepts they represent;
- Allow for a larger entry vocabulary, which would directly help both finding the standard reference class fast and improve recall in automated classification.

Further research should include determining other reasons behind browsing failures in the Ei classification scheme. It should also deal with problems of disparate evaluations of one and the same web page. This could mean applying a different user study methodology that would help participants make more homogeneous decisions. Another way would be to harvest web pages that are more uniform in terms of quality or genre (see, for example, Custard and Sumner, 2005; Nicholson, 2003), or to design more narrowly specified search tasks for a certain purpose.

Acknowledgments

We thank Birger Larsen who provided suggestions on the user study design, based on the interactive track at INEX2005. We also thank Anders Ardö and Traugott Koch who provided useful comments on earlier versions, significantly improving the paper. This work was supported by the IST Programme of the European Community under ALVIS (IST-002068-STP).

References

Aitchison, T.M., and Harding, P. (1982), "Automatic indexing and classification for mechanized information retrieval", *Proceedings of EURIM 5, 12.-14.5.1982, Versailles.*

Ardö, A. (2007), "Focused crawler: Combine system homepage", available at: http://combine.it.lth.se/ (accessed 3 September 2007).

Borlund, P. (2003), "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems", *Information Research*, Vol. 8 No. 3, paper no. 152, available at: http://informationr.net/ir/8-3/paper152.html (accessed 5 September 2007).

Custard, M. and Sumner, T. (2005), "Using machine learning to support quality judgments", *D-Lib Magazine*, October 2005, Vol. 11 No. 10, available at: http://www.dlib.org/dlib/october05/custard/10custard.html (accessed 3 September 2007).

DESIRE (2000), "DESIRE: Development of a European Service for Information on Research and Education", DESIRE, available at: http://www.desire.org/ (accessed 13 February 20049).

Engineering Information (2006), "Compendex", Engineering Information, Elsevier, available at: http://www.ei.org/databases/compendex.html (accessed 30 June 2006).

Golub, K. (2006a), "Automated subject classification of textual Web documents", *Journal of Documentation*, Vol. 62 No. 3, pp. 350-371.

Golub, K. (2006b), "The role of different thesauri terms in automated subject classification of text", *Proceedings of the International Conference on Web Intelligence, Hong Kong*, pp. 961-965.

Golub, K., and Ardö, A. (2005), "Importance of HTML structural elements and metadata in automated subject classification", *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries, Vienna, Austria, 18-23 September*, pp. 368-378.

Golub, K., Hamon, T., and Ardö, A. (2007), "Automated classification of textual documents based on a controlled vocabulary in engineering", *Knowledge Organization*, Vol. 34 No 4, pp. 247-263.

Ingwersen, P., and Järvelin, K. (2005), *The turn: Integration of information seeking and retrieval in context*, Springer, Dordrecht.

Intute Consortium (2006), Intute: Science, engineering and technology – engineering general, available at: http://www.intute.ac.uk/sciences/cgi-bin/browse.pl?id=25682 (accessed 30 August 2007).

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.

Koch, T., and Zettergren, A.-S. (1999), "Provide browsing in subject gateways using classification schemes", EU Project DESIRE II, available at http://www.mpdl.mpg.de/staff/tkoch/publ/class.html.

Koch, T., Golub, K., and Ardö, A. (2006), "Users browsing behaviour in a DDC-based Web service: a log analysis", *Cataloging & Classification Quarterly*, Vol. 42 No. 3/4, pp. 163-186.

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

Larsen, B., Malik, S. and Tombros, A. (2006), "The interactive track at INEX2005", *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005, Revised Selected Papers*, pp. 398-410.

Lewis, C. and Rieman, J. (1994), "Task-centered user interface design: a practical introduction", available at: http://www.hcibib.org/tcuid/ (accessed 3 September 2007).

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/.

Luhn, H. P. (1957), "A statistical approach to mechanized encoding and searching of literary information", *IBM Journal of Research and Development*, 1(4), pp. 309-317.

McMahon, C. et al. (2004), "Waypoint: An Integrated Search and Retrieval System for Engineering Documents", *Journal of Computing and Information Science in Engineering*, Vol. 4, DOI: 10.1115/1.1812557, available at http://www.icbl.hw.ac.uk/~santiago/GoldDust/docs/WayPoint.pdf

Milstead, J, ed. (1995), *Ei thesaurus*, 2nd ed., Engineering Information Inc., Hoboken, NJ.

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

Nicholson, S. (2003), "Bibliomining for automated collection development in a digital library setting: using data mining to discover Web-based scholarly research works", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 12, pp. 1081-1090.

Nielsen, M.L. 2004, "Task-based evaluation of associative thesaurus in real-life environment", *Proceedings of the ASIST 2004 Annual Meeting, Providence, Rhode Island, November 13-18*, pp. 437-447.

Nübel, R. et al. (2002), "Bilingual indexing for information retrieval with AUTINDEX", *Third International Conference on Language Resources and Evaluation, 29th, 30th & 31st May, Las Palmas de Gran Canaria (Spain)*, pp. 1136-1149.

Plaunt, C., and Norgard, B.A. (1998), "An association-based method for automatic indexing with controlled vocabulary", *Journal of the American Society for Information Science*, Vol. 49 No. 10, pp. 887-902.

Salton, G. (1989), *Automatic text processing: the transformation, analysis, and retrieval of information by computer*, Reading, MA: Addison-Wesley.

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

Sheskin, D.J. (2000), *Handbook of parametric and nonparametric statistical procedures*, 2nd ed., Chapman & Hall, Boca Raton etc.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol. 34 No. 1, pp. 1-47.

Soergel, D. et al. (2004), "Reengineering thesauri for new applications: the AGROVOC example", *Journal of Digital Information*, Vol. 4 No. 4, Article no. 257, available at: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/.

Vaughan, L. (2001), *Statistical methods for the information professional: a practical, painless approach to understanding, using, and interpreting statistics*, Information Today, Inc., Medford, NJ.

Vizine-Goetz, D. (1996), "Using library classification schemes for internet resources", OCLC Internet Cataloging Project Colloquium, available at: http://webdoc.sub.gwdg.de/ebook/aw/oclc/man/colloq/v-g.htm (accessed 4 April 2006).