Browsing engineering resources on the Web: a general knowledge organization scheme (Dewey) vs. a special scheme (EI)

Abstract: Under the auspices of the Desire II project, researchers at NetLab and OCLC are providing searching and browsing of a test collection of engineering documents on the Web. The goal of the project is to explore simple methods of automatic classification to provide subject browsing of a robot-generated engineering index. At NetLab the documents are automatically classified and organized using an engineering-specific scheme, the Engineering Index (Ei) Thesaurus and Classification; at OCLC the Dewey Decimal Classification (DDC), a general knowledge organization scheme, is being used.

1. Introduction

The goal of the DESIRE II project is to explore automated methods for gathering and organizing Web resources to improve resource discovery on the Internet. Manual efforts, even on a large scale, are insufficient to keep up with the quantity of changes to documents in a service or subject area of some size. As part of DESIRE I, NetLab researchers developed an approach for integrating a manually selected and cataloged collection of WWW-resources (the Engineering Electronic Library, Sweden or EELS) with a much larger robot-generated subject index in the same subject area. In DESIRE II, NetLab researchers tested and evaluated different methods of automatic classification based on the Engineering Information (Ei) scheme. The automatic classification and cross-browsing functionality will be added to the EELS service. OCLC researchers are studying additional searching and browsing options that might result from the use of the Dewey Decimal Classification (DDC), a general knowledge organization scheme, to organize the same collection of resources.

2. Background

In 1994, NetLab developed the EELS service on behalf of the Swedish Universities of Technology Libraries. In the EELS service, Internet resources about engineering (about 1,400) are selected according to agreed-upon quality criteria, described and supplied with thesaurus terms and classification codes based on the Engineering Information (Ei) scheme. In 1996, the robot-generated subject index, “All” Engineering (AE), was added to the EELS service. The experimental full-text index of international engineering resources was implemented as a complement to the smaller database of quality-assessed EELS resources. The AE index contains 127,192 resources and is searchable as full text, and by title, headings, URL, anchor text, and outgoing links. Additionally, AE can be browsed alphabetically by resource title, by domain or country, or by the number of times a resource has been linked to (i.e., cited by) other pages in the database. As a DESIRE I project development, the manually created EELS service and the robot-generated AE index were integrated and offered as an experimental service. A simultaneous search in both databases is made possible by using the Z39.50 search protocol and the Dublin Core metadata standard. The system differentiates
between the two levels of quality by displaying the hits from the EELS first in the results display. From the list of results, a user can seamlessly navigate and search in both services.

As implemented in DESIRE I, searching across the two services works well, but browsing by Ei subject classification is possible only in EELS not in “All” Engineering. AE does not have a corresponding subject structure that will support this function. At the end of DESIRE I, it was deemed highly desirable to offer browsing from the small collection of quality-controlled resources in EELS to equivalent resources in the large index; however, given the size of the robot-generated engineering index (nearly 130,000 documents, in the AE index) it become clear that automated classification techniques would be needed to provide a subject-based browsing interface to the index.

3. Test collection of engineering resources

In addition to investigating automated classification techniques in DESIRE II, NetLab researchers also explored improvements to the technique used to generate the engineering index. Ardo, Koch, and Noodén (1999) explored various methods for gathering an engineering collection, and creating a stable database from the resources. Seven approaches for building the robot-generated index were considered. The following two were investigated in detail. In approach 1, the harvesting and indexing software was started at the top page of a small number of engineering link collections (30), and all links were followed in three steps resulting in a database of 215,589 documents. In approach 2, seven quality-controlled engineering collections were harvested completely, and all links were followed in two steps resulting in a database of 159,172 documents. Both collections had about the same percentage of relevant documents, 77%, although the overlap between the two databases was small. Only 18% of URLs overlap when the two databases are combined. The overlapping URLs account for 27% of the resources in database 1, and 36% of the resources in database 2. Although the two approaches produce similar results with respect to relevance and uniqueness, the linking behavior in the last step has a large influence on the composition of the databases and accounts for about 90% of the resulting resources.

For a real service, like NetLab’s AE service, these findings indicate that it is necessary to combine both approaches to produce a database with sufficient subject coverage in engineering. For the purposes of investigating automated classification methods in the DESIRE II project, we decided that approach 2 produced an adequate test database. Of the approximately 156,000 usable documents harvested, 132,000 were in English. Only English-language documents were subject to automatic classification in the later phases of the project, since both the Ei scheme and DDC edition used are in English.

4. Automatic Classification in Desire II at NetLab

To provide a subject based browsing interface for the test collection, NetLab researchers generated an Ei-based classification structure equivalent to the one used in the EELS service. The Ei Thesaurus, which contains more than 16,000 terms intellectually mapped to more than 800 classification categories, was used to provide a rough subject classification of the resources.

For each document in the test database, all metadata, headings, and plain text were extracted. Each of the vocabulary terms from the thesaurus, and captions from the classification categories, was matched against this text. When a match was found, the corresponding classification codes were associated with the record. A score was also assigned to each code based on term complexity (single word, Boolean expression or phrase), type of classification (main or optional), match location (metadata, headings or plain text), and match frequency. All scores were summed for each class. Scores from classes directly above in the Ei hierarchy structure were added to the scores for the most specific classes.
below in order to assign the most specific classification to a document. For every document, a list of classification codes in decreasing order by score was generated.

Various heuristics, weighting schemes, and cut-off points were applied in an effort to improve the classification process, and the resulting browsing structure (Ardó & Koch, 1999; Koch & Ardó, 2000). The results were achieved without using word stemming. Instead, an expanded stop word list, and normalized weights with relative matching frequencies, was used. A cut-off threshold of 3% of the total page scores, and an absolute score of 2, determines which documents are displayed in a given class in the browsing system. The average number of classifications per page is slightly less than 5. Since stemming was not used, 11% of the documents do not receive a classification code. With stemming, 98% of the documents would be classified, but at the expense of many erroneous classifications.

The outcome of the automated classification process was evaluated on the following aspects: 1) the ability of the system to correctly classify the Ei classes themselves; 2) the distribution of documents across the classification scheme; and 3) the comparison of the automatic classification against intellectual classifications. Regarding the first, the automated classification system correctly classified the Ei classes with 100% accuracy. The distribution of documents by Ei main class and level of hierarchy is shown in Table 1.

<table>
<thead>
<tr>
<th>Main class/Level</th>
<th>Civil (4)</th>
<th>Mining (5)</th>
<th>Mechanical (6)</th>
<th>Electrical (7)</th>
<th>Chemical (8)</th>
<th>Engineering, General (9)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>598</td>
<td>598</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>5</td>
<td>18</td>
<td>0</td>
<td>19</td>
<td>113</td>
<td>155</td>
</tr>
<tr>
<td>3</td>
<td>28,224</td>
<td>10,472</td>
<td>14,120</td>
<td>27,454</td>
<td>10,247</td>
<td>17,272</td>
<td>107,789</td>
</tr>
<tr>
<td>4</td>
<td>112,081</td>
<td>25,753</td>
<td>40,723</td>
<td>127,998</td>
<td>42,495</td>
<td>13,0024</td>
<td>479,074</td>
</tr>
<tr>
<td>5</td>
<td>3,230</td>
<td>15,743</td>
<td>3,316</td>
<td>4,451</td>
<td>1,796</td>
<td>9,254</td>
<td>37,790</td>
</tr>
<tr>
<td>Total</td>
<td>143,535</td>
<td>51,973</td>
<td>58,177</td>
<td>159,903</td>
<td>54,557</td>
<td>157,261</td>
<td>625,406</td>
</tr>
</tbody>
</table>

Table 1: Distribution of documents by Ei Main Class and Level of Hierarchy

The distribution of class mappings to terms mirrors the distribution of documents at that level. For example, the two top levels are practically empty because there are no thesaurus terms mapped to these classes by the Ei editors (the Ei class captions result in the few assignments shown). The classification process assigns a slightly higher percentage of the documents to level three than would be expected based only on the number of terms at this level.

The distribution across main subject categories is different. Main classes (7) Electrical, and (9) Engineering, General are the largest, followed closely by (4) Civil. The three remaining classes, (6) Mechanical, (8) Chemical, and (5) Mining, are of roughly equal size but only about one third the size of the largest classes. Class 9 contains 25% of the classifications, and 12% of the terms, while main class 5 has only 8% of the classifications, but about 14% of the terms. General engineering is better represented among the real documents on the Web than mining engineering is. This uneven distribution of documents by subject on the Web accounts for the uneven distribution of documents over the Ei classification categories and thesaurus terms.

Last, intellectually assigned classifications for about 1,000 Web pages from the EELS service were compared with the automatic classification for the same pages. The comparison showed an identical or more specific automatic classification in 57-66% of the cases. This comparably good result is most likely due to the richness of the Ei vocabulary system, which contains a significant number of terms that have been intellectually mapped to the Ei classes.

5. Automatic Classification at OCLC

At OCLC, the Dewey Decimal Classification (DDC) is being used to provide a subject structure for the test collection of engineering resources. The DDC was selected for use
because it is published by Forest Press, a division of OCLC, and it is the most widely used general classification system in the world. DDC has well-developed hierarchies, well-defined categories, a rich network of relationships, and an enriched vocabulary base. DDC's notation and structure provide context and relationships for topics, a feature that has great potential for improving browsing and retrieval in the electronic environment. The version of the DDC that is used in the project is the Enhanced DDC database.

The Enhanced DDC database contains the latest version of the DDC editorial database for the full edition of the DDC. The DDC editorial group updates and maintains the DDC using an editorial system developed at OCLC. The Enhanced DDC database includes several mechanisms for incorporating new terminology. It contains nearly 5,000 more index terms than in the print edition. Also included are more than 5,400 Library of Congress subject headings (LCSH) that have been mapped to the DDC by the DDC editorial staff. The editorial mappings include new LCSH selected from the LC Weekly Lists of Subject Headings. The LCSH from the Weekly Lists are often a source of new technical terms: “Image analysis”, “Microelectromechanical systems”, “Mixed signal circuits”, “Nanowires”, and “Solar gas turbines”, are examples of technical terminology added to DDC from the Weekly Lists. Layered on to the additional index terms and LCSH mappings, are an additional 100,000 subject terms from several sources that have been mapped to the DDC using a variety of techniques (Vizine-Goetz, forthcoming).

A program called “Scorpion” (Shafer, n.d.) is used to do the automatic class number assignment. The automated classification database is built from the Enhanced DDC database. During the automated classification process, a set of terms is extracted from the resource and used to retrieve a ranked list of DDC numbers from the terminology-enhanced database. The ranking is based on relative frequency of occurrence of terms and other criteria relating to associations found in the DDC classification. The highest ranking class numbers are then associated with the resource.

The WordSmith software developed at OCLC identifies terminology in raw text using a variety of computational linguistics methods (Godby & Reighart 1999). One heuristic works on noun phrases and computes a weighted frequency based on the frequencies of grammatically significant information, such as the head, or the rightmost word, of the noun phrase. For example, since "engineering" is a highly posted head noun in the EELs data, noun phrases containing the term "engineering," such as "biomedical engineering," are weighted more highly than would be warranted by their raw frequencies alone. One outcome of the WordSmith processes is a relatively small set of high-quality topical vocabulary that would be suitable as an index or browse display and can supplement the subject indexes provided by the Ei Thesaurus or the DDC. The same processes can be used in automatic classification. Our hypothesis is that, the precision of automatic classification can be improved if the input to the classifier is represented as a list of the most significant noun phrases instead of the complete text of the raw document. The classifications reported in this study are obtained by creating surrogate documents that consist of noun phrases whose score on two heuristics that measure term stability and topicality exceeded the means for the collection. For example, a raw document containing about 2,300 terms is reduced to the following set of terms by the WordSmith processes:

javaworld
java collections framework
advertising information
java
programming java threads

Automatic classification of the WordSmith document and the corresponding raw document produce the results shown in Table 2.
In this example, there is little difference among the top five ranked classes assigned by each approach. Both assign DDC classes 005.7, 005.758, 005.71376 as well as an additional class number from the DDC hierarchy, 005.71 (Data communications). Both assign a number that is slightly off the mark--DDC number 070.579 from the hierarchy for publishing (WordSmith) and 025.04 for information storage and retrieval systems in general (raw document). According to the DDC, computer science aspects of information storage and retrieval systems should be assigned class number 005.74.

Although the two approaches appear to produce similar assignments for a given document, analysis of the aggregated results produces a different view. When WordSmith classifications, and raw document classifications are compared by DDC main class, WordSmith produces a more even distribution of class number assignments (see Table 3). This distribution of DDC topics may better represent the topics in the actual documents than the raw document-based distribution of DDC classes. This hypothesis will be tested by performing intellectual assessments of the accuracy of class assignments for each approach.

Both approaches assign roughly the same percentage of class numbers from DDC main classes likely to contain engineering-related topics (000, 300, 500, and 600) with 91.1% using WordSmith and 92.9% using the raw documents. Additionally, both approaches appear to assign DDC numbers at about the same level of specificity (see Tables 4 and 5).
### Table 5. Raw Document Classifications by Level of Specificity

<table>
<thead>
<tr>
<th>Level</th>
<th>000</th>
<th>300</th>
<th>500</th>
<th>600</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.8</td>
<td>1.3</td>
<td>4.1</td>
<td>1.4</td>
<td>2.2%</td>
</tr>
<tr>
<td>2</td>
<td>8.4</td>
<td>5.3</td>
<td>14.9</td>
<td>4.3</td>
<td>7.3%</td>
</tr>
<tr>
<td>3</td>
<td>24.3</td>
<td>14.8</td>
<td>30.3</td>
<td>17.4</td>
<td>20.4%</td>
</tr>
<tr>
<td>4</td>
<td>32.3</td>
<td>22.9</td>
<td>32.3</td>
<td>34.7</td>
<td>29.0%</td>
</tr>
<tr>
<td>5</td>
<td>17.4</td>
<td>33.4</td>
<td>12.5</td>
<td>20.8</td>
<td>18.8%</td>
</tr>
<tr>
<td>6</td>
<td>14.9</td>
<td>22.2</td>
<td>5.9</td>
<td>21.3</td>
<td>15.2%</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>92.9%</td>
</tr>
</tbody>
</table>

#### 6. Next Steps

The previous discussion indicates that additional research is needed to develop optimal methods of automated classification using a general knowledge organization scheme like Dewey. OCLC researchers are continuing to refine their techniques using the stable, engineering test collection created by their NetLab colleagues. The results will be evaluated in a manner similar to NetLab's so that a comparison of each group's automatic classification approaches can be made. At the conclusion of the project, OCLC will provide a DDC-based system for searching and browsing the test collection of engineering resources.

#### Notes

1. Desire is a major international project aiming to build large-scale information networks for the European research community. See DESIRE: Development of a European Service for Information on Research and Education. EU project. [http://www.lub.lu.se/desire/](http://www.lub.lu.se/desire/) Accessed 03/16/00. Background information on the Desire II project NetLab/OCLC collaboration is available in the Annual Review of OCLC research. Accessed 03/16/00.


#### References


