Libri, 2002, vol. 52, pp. 48–53
Printed in Germany · All rights reserved

Libri
ISSN 0024-2667

Innovative Solutions in Automatic Classification: A Brief Summary

ERZSÉBET TÓTH
Department of Library and Information Science, Institute of Mathematics and Informatics, College of Nyíregyháza, Hungary

There is a growing need for practical solutions to provide flexible access to digital documents in a structured form on the Web. The existing library classification schemes serve as good bases for achieving this goal. This paper presents a brief review of the various methods applied in automatic classification. It focuses on the main activities fulfilled within various research projects to make possible the effective automatic indexing and classification of Web sources. It describes the approaches taken in the Nordic WAIS/WWW; DESIRE II – Engineering Electronic Library System (EELS); GERHARD; and SCORPION projects. Artificial neural networks and artificial intelligence show great potential.

The current state of research projects

The experimental projects involved in automatic classification are considered to be prominent research activity in the field of library classification. Previously automatic classification was based on various clustering and statistical methods. However, concrete implementations were hampered by problems arising from the lack of computerisation and limited storage. Nowadays there are a growing number of research projects that employ the former indexing techniques and also examine the efficiency of different clustering methods.

At present there do not appear to be any practical examples where library classification systems have been completely overtaken by automatic methods. However there is an increasing interest in developing such systems. In most cases multinational corporations support the research projects. This ambition on the one hand can be explained by the need to provide a practical solution where digital documents are easily accessible in a structured form on the Web. This effort on the other hand is due to the fact that the simple indexing techniques of Internet search services do not provide adequate results in terms of precision and relevance.

In general the Universal Decimal Classification (UDC) is a useful tool for organising Web resources since its machine-readable format is widely spread among Internet developers. A more rigorous approach to defining the content of Web documents can be detected at those Web sites that integrate UDC into their operation. This type of classification contrasts with the practice of those services that use DDC and the Library of Congress Classification Scheme, since they are characterised by a very simplified notation and a lack of details (Newton 2000).

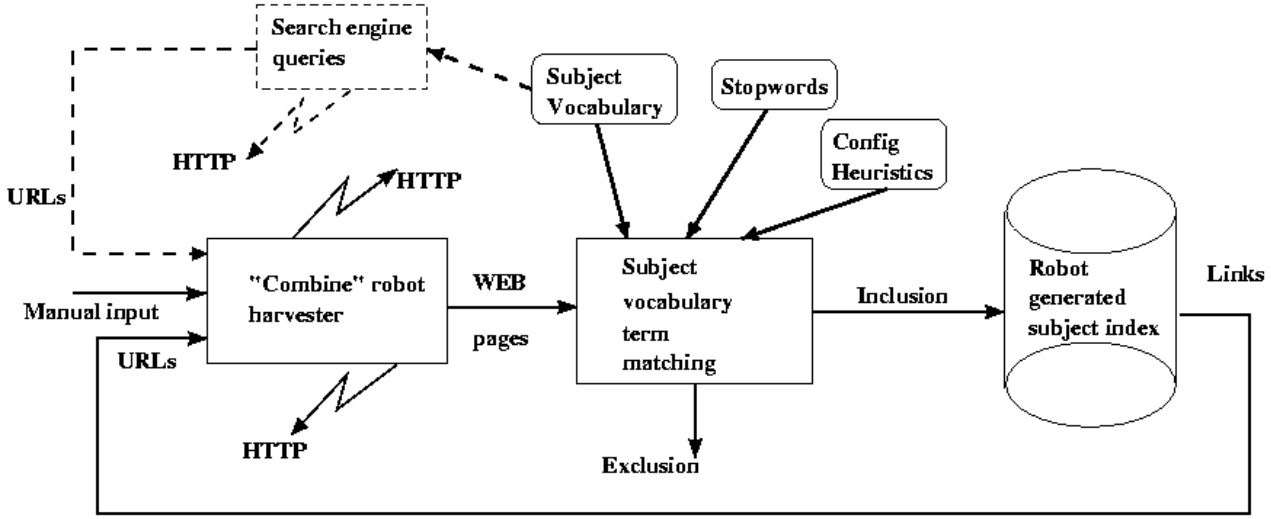
Nordic WAIS/World Wide Web project

In the summer of 1993 the Lund University Library and The National Technological Library of Denmark launched the Nordic WAIS/World Wide Web project. It lasted one year. It has accomplished the automatic classification of 660 WAIS databases, but it was limited to 51 UDC classification categories. It has exploited the machine-readable format of the UDC to a great extent (Ardö 1999a, Newton 2000).

Its technique was the following. Keyword lists were generated for each WAIS database. In this case keywords were extracted from the subject

Erzsébet Tóth, Könyvtár és Alkalmazott Informatikai Csoport, Matematika és Informatika Intézet, Nyíregházi Főiskola, Sóskói ut., 31/b., 4400 Nyíregyháza, Hungary. E-mail: tothe@zeus.nyf.hu

Figure 1. Retrieving a robot-generated subject index with a topical filter



fields of the database descriptions. Then these keywords were matched with the classification categories of the UDC scheme. If an exact match was found classification codes were generated as a result. Codes were weighted according to the field of the database description from which the keywords came. Finally these weightings were compared and the most appropriate classification codes were selected.

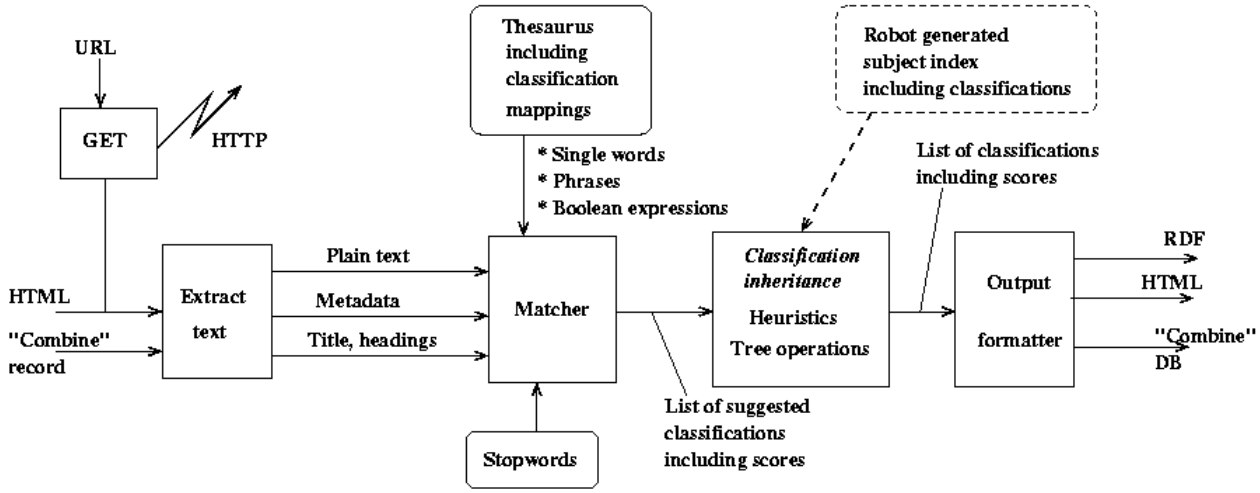
DESIRE II project

The general purpose of the European Union's DESIRE project (Development of a European Service for Information on Research and Education) is to enhance the established European information services for the research community. The second part of the project started in July 1998 with the participation of ten partners, and lasted two years (DESIRE 1999). They analysed how it was possible to integrate a manually selected link collection with a much larger robot-generated subject index. The research was based on the Engineering Electronic Library System (EELS) subject gateway and the robot-generated subject index called 'All Engineering'. They generated the browsing interface of the subject index by means of the Engineering Information Inc. (EI) structure, which had previously been used in the EELS service.

First of all they had to solve the problem of retrieving relevant items for the subject index. They managed to accomplish this task with the use of the 'Combine' robot harvester. In their retrieval model they used a simple matching algorithm with a thesaurus to filter Web pages. This matcher compared the text of sources with the terms of a subject-specific vocabulary. They applied various heuristics and weighting schemes to the results to determine whether a page should be retrieved to the index database. It meant that certain Web pages would be excluded from the index database if their texts did not correspond with the terms of the subject vocabulary. This process is well illustrated with the 'inclusion' and 'exclusion' arrows in Figure 1. This retrieval process does not simply end at the subject index because a new input should be entered onto the index database in the form of relevant Web pages. To carry out the subject classification of Web pages they used the EI thesaurus in which terms were intellectually mapped to EI classification categories.

Initially metadata was collected for the index database from the HTML (Hypertext Mark-up Language) meta tags, headings and body of the resources. Then the generated index terms were matched with all thesaurus terms, taking stopwords into account. If the algorithm found a match, a list of class codes was associated with the document. The index terms and the related class codes in a document were weighted. The reason for this weighting was to differentiate the relevance of the index terms since phrases and Boolean terms describe the semantic content of a document more precisely than single-word terms.

Figure 2. Automatic classification process



Finally a list of suggested classifications for each document was generated with scores in decreasing order. To select the relevant classifications for display in the service they utilised some heuristic post-processing. The number of the classification codes was reduced with truncation. So it meant that all classifications were ignored whose absolute value was lower than a threshold. The value of the threshold was determined with a heuristic. The relevant classification codes of the service were displayed in RDF (Resource Description Framework – http://www.w3.org/RDF/) and HTML formats. Figure 2 illustrates this process.

The results of this experiment were as follows: they applied automatic classification to 86,468 documents all together, and 6 classification codes were assigned to a Web page on average. They analysed 923 Web pages as a sample and found that the proportion of automatically generated classification codes corresponding to the ones achieved by intellectual classification was between 57 and 66% (Ardö 1999b).

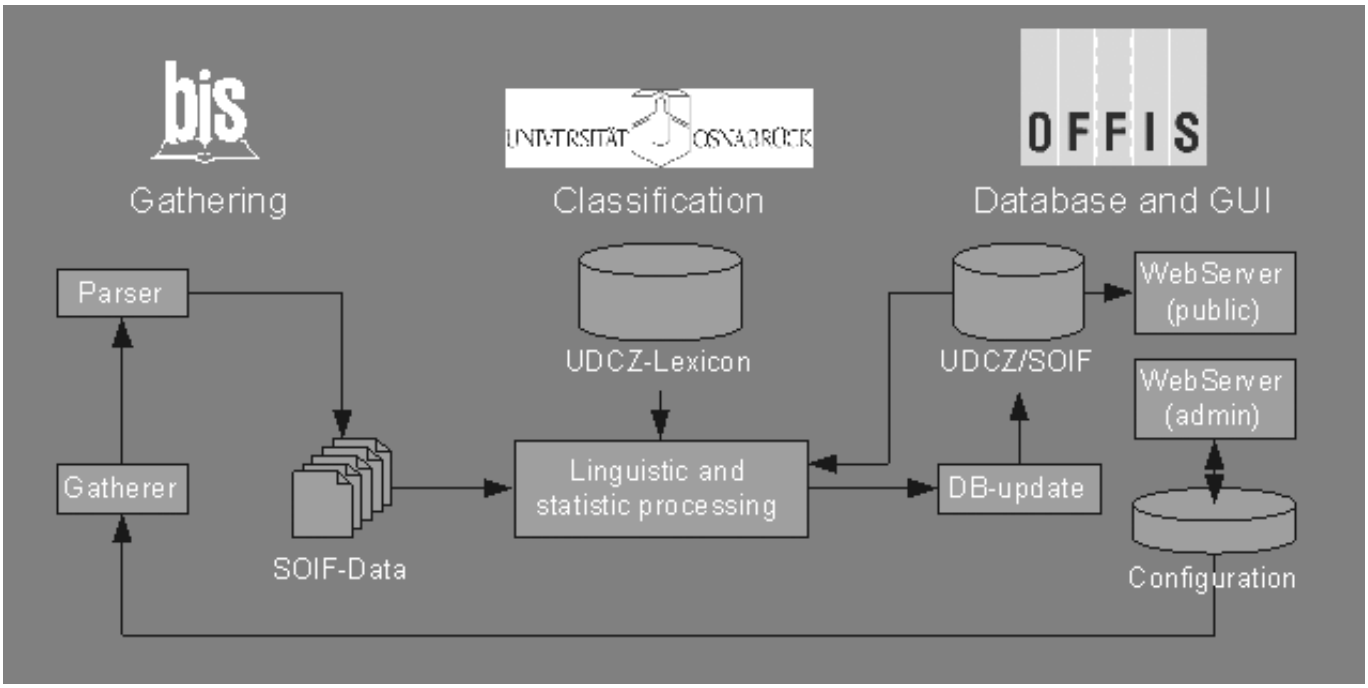
GERHARD project

The GERHARD (German Harvest Automated Retrieval and Directory – http://www.gerhard.de/) system was focused on the automatic indexing and classification of German Web pages. Its service integrates searching and browsing facilities for users. The developers applied a trilingual version of UDC (Universal Decimal Classification), updated according to ETH-Zürich and called UDCZ. This scheme comprises 60.000 classification categories (entries). Fifteen different relations are possible among them. Each entry includes a UDC code (notation), a category description and synonyms in German, English and French.

A complex system architecture may be observed in Figure 3 comprising different parts such as: retrieval, parsing, Summary Object Interchange Format (SOIF) data, linguistic and statistical processing which occurs with the help of the UDCZ lexicon. At the end of this process UDCZ codes and SOIF formats were entered onto an Oracle relational database. In GERHARD the robot harvester collected the relevant pages to an index database by using filtering rules. To determine the exact place of relevant Web pages they stored all configuration data in a relational database that could also be accessed and modified through a Web interface. After retrieving data, another software parsed the documents thoroughly and prepared them for further processing. At last the essential content of a document was stored in the structured format called Summary Object Interchange Format.

Two basic demands have influenced the implementation of automatic classification, namely 'maximum quality' and 'minimum time consumption'. Different linguistic and statistical methods were applied to achieve these purposes. Their main idea was to convert the UDCZ scheme to a lexicon that mapped classification entries to UDCZ codes. They matched the natural language expressions of the documents with the entries of the UDCZ lexicon by means of linguistic classifi-

Figure 3. System architecture



cation. If the two phrases corresponded with each other a list of classification codes was generated for the document. Later on codes were weighted in order to find and display the most precise and relevant ones. They reduced the number of the notations by using various statistical methods. After statistical post-processing the SOIF formats and the classification codes were held in the tables of the Oracle relational database.

On the basis of Figure 4 it can be said that a linguistic classification consists of three basic components: UDCZ conversion, UDCZ lexicon construction; text conversion and analysis; UDCZ notation analysis and selection. The input for text conversion and analysis are documents with ASCII texts and the final result will be texts with the selected notations.

The primary aim of the UDCZ conversion was to extract those natural language expressions from the raw data of the UDCZ scheme that described classification categories. It demanded the elaborate processing of the linguistic information available in the UDCZ.

The UDCZ conversion was fulfilled in the following three steps:

1. Words in the UDCZ categories were analysed from a morphological aspect and they were reduced to their stems. Their word class information appeared in an annotation next to them. Various linguistic software was used for this purpose e.g. Lingsoft's GERTWOL and ENGTWOL programs.
2. Different rules were applied to extract and construct natural language expressions from the analysed texts.
3. Abbreviations, stop words and annotations were eliminated from the category descriptions.

The automatic process of the UDCZ conversion is illustrated with the following three tables:

In the raw data of the UDCZ scheme each entry includes a UDCZ code and the natural language description of the UDCZ category in three languages (see Table 1). In the first two lines of Table 2 stems and tagged word forms can be seen. In the last two lines of the same table the generated natural language expressions are pre-

Table 1: Unstructured data entry of the UDCZ

```
001Z~03
002DDUEBERSETZUNGEN/TECHNISCHE U. NATURWISSENSCHAFTLICHE
003DETRANSLATIONS/TECHNICAL AND SCIENTIFIC
004DFTRADUCTION/SCIENTIFIQUE ET TECHNIQUE
```

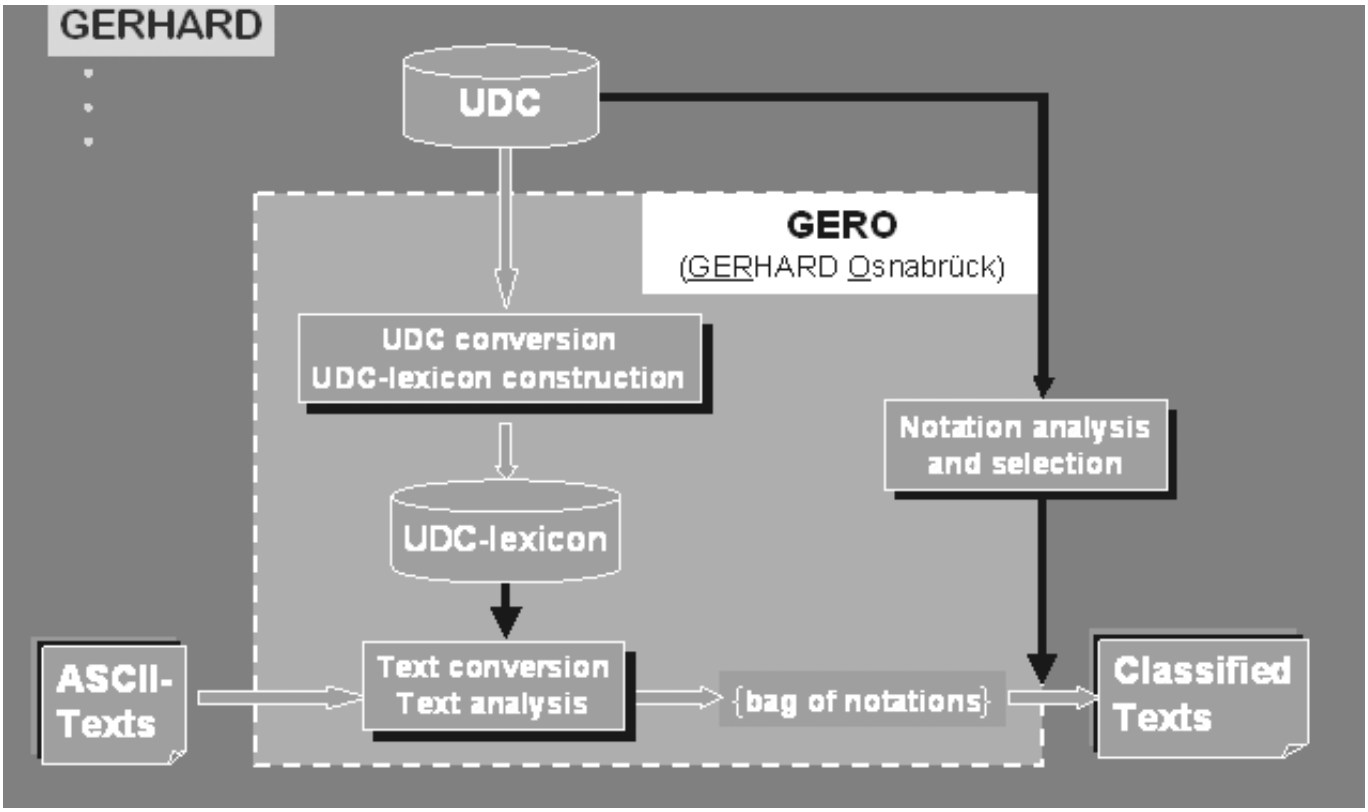
Table 2: Results of the UDCZ conversion

```
translation~~N/technical~~Adj and~~Conj scientific~~Adj
uebersetzung~~N/technisch~~Adj u.~~Conj naturwissenschaftlich~~Adj
technical translation; scientific translation
technisch uebersetzung; naturwissenschaftlich uebersetzung
```

Table 3: UDCZ-lexicon categories

```
technical translation:- :~03
gene:xxx s:575.113.1
```

Figure 4. Linguistic classification used in GERHARD



sented. Table 3 shows classification categories with their associated codes.

The next logical step is text conversion and analysis, which can be summarised in this way: the text of the documents had to be adjusted to the formal requirements of the UDCZ lexicon (e.g. deletion of stop words). Software (recogniser) was constructed from the UDCZ lexicon which matched strings in the document text with the classification categories. The essence of its operation was the following: the software added various truncation variables to word stems to support an exact match with specific word forms during text classification (e.g. '*technische uebersetzungen*', '*technischer uebersetzungsvorschriften*'). These variables are useful in general because they provide flexible textual matches. However they result in false matches in the case of short words for example 'gene' would correspond to 'general' and 'generic'. Therefore they collected all possible morphological endings of a word to make a difference between them. In Table 3 '-' signifies arbitrary endings and 'xxx' means that the stem is the same as the word form.

If the software found an exact match, a list of classification codes was associated with the document. The selection of relevant codes was done in two steps:

1. Information given in the notations was exploited and their frequency of occurrence was also analysed. On the basis of their textual match those clusters of notations were found that they belonged to.
2. The generated notations were weighted and reduced by means of statistical and heuristic processing. Relevant notations were filtered with the use of appropriate algorithms.

In fact the navigation is supported by a directed graph, which includes separately classification categories and relations among them. It consists of several cycles having categories on its nodes and implicit relations on its edges. In the graph it is possible to store in separate groups: the duplicate records, cross-references to earlier used classification codes, new categories and starting points for navigation (Möller 1999).

SCORPION project

The SCORPION research project (http://orc.rsch.oclc.org:6109/) was launched by Online Computer Library Center (OCLC) in 1998. It examines the

different methods of automatic classification and connects indexing with cataloguing. It is similar to GERHARD in the sense that it also uses linguistic and statistical methods (Möller 1999). SCORPION software is built on a searchable database constructed from the data files of Dewey Decimal Classification. In this database adequate classification categories are mapped to Dewey codes (Hickey 1999). Here a document can be regarded as a query against a database. As a result of the retrieval Dewey codes are assigned to a document in a ranked way (Maclennan 2000).

Keywords are automatically extracted from the documents that are also taken into consideration in the retrieval and ranking of DDC numbers (Hickey 1999). They apply a weighting scheme to DDC codes to count the number of occurrences of keywords in the database records. They use cosine normalisation for computing the angle between the vector representations of a record and a query (Shafer 1997).

Possible trends in software development

This review is intended to be a short summary of the advances that have been made in this particular field. It can be seen that several approaches were utilised in these projects, including heuristics, weighting schemes, and computer linguistic methods. A close cooperation among the projects' researchers came to exist which helped them to refine their research methods. These research projects are promising and their findings might be useful for investigating other alternative solutions for Internet services. There is a growing need to change and adapt library classification systems to become a suitable browsing tool for these services. Flexible browsing requires the development of tailor-made visualisation and navigation techniques. The focus seems likely to be on clustering methods, content-based, usage pattern or citation-based techniques. There is great potential in artificial neural network and AI techniques, although they have been previously used in other fields (Ardö 1999b).

Editorial history:
final version received 18 February 2002;
accepted 25 February 2002.

References

Ardö, A. *et al.* 1999a. Improving resource discovery and retrieval on the Internet: the Nordic WAIS/ World Wide Web Project – Summary Report. *Nordinfo-NYTT* 17 (4): 13–28.

Ardö, A., and Koch, T. 1999b. Automatic classification applied to full-text Internet documents in a robot-generated subject index. *In*: Online Information 1999: the proceedings of the 23rd International Online Information Meeting. London, 7–9 December, pp. 239–246. URL: http://www.lub.lu.se/~anders/online99 [viewed 28 February 2002].

DESIRE (Development of a European Service for Information on Research and Education). 1999. EU Project. URL: http://www.lub.lu.se/desire [viewed 28 February 2002].

Hickey, T. and Vizine-Goetz, D. 1999. The role of classification in CORC. In: Online Information 1999: the proceedings of the 23rd International Online Information Meeting. London, 7–9 December, pp. 247–250.

MacLennan, A. 2000. Classification and the Internet. *In*: Marcella, R., and Maltby, A.: The future of classification. Aldershot/Brookfield, Vt.: Gower, pp. 59–68.

Möller, G. *et al*. 1999. Automatic classification of the World Wide Web using Universal Decimal Classification. *In*: Online Information 1999: the proceedings of the 23rd International Online Information Meeting. London, 7–9 December, pp. 231–237.

Newton, R. 2000. Information technology and new directions. *In*: Marcella, R., and Maltby, A.: The future of classification. Aldershot/Brookfield, Vt.: Gower, pp. 43-57.

Shafer, K., and Thompson, R. 1997. Scorpion: SMART Weighting Schemes. URL: http://orc.rsch.oclc.org:6109/smart_weight.html [viewed 28 February 2002].