Different Approaches to Automated Classification: Is There an Exchange of Ideas?[1]

Koraljka Golub*, Birger Larsen**

*koraljka.golub@it.lth.se
KnowLib, Department of Information Technology, Lund Univeristy, P.O. Box 118, SE-221 00 Lund (Sweden)

**blar@db.dk
Department of Information Studies, Royal School of Library and Information Science,
Birketinget 6, DK-2300 Copenhagen (Denmark)

Abstract. Automated classification of text has been studied by three major research communities, machine learning, information retrieval, and library science, each taking a different approach. The paper aims to study to what a degree the three communities explore others' ideas, methods, findings. To that purpose we studied direct links (do authors from one community cite authors from another) and indirect links (using bibliographic coupling). Although the study is based on a small sample of 148 papers, the results indicate that the three communities do not exchange ideas to a great extent.

Introduction

Automated subject classification has been a challenging research issue for several decades. The interest has grown rapidly with the emergence of the World Wide Web (WWW) and related digital information services with very large amounts of documents, where the high costs of manual subject classification is a major hindrance.

Currently, there are three distinguishable approaches to automated subject classification of text, each taken by a different research community: text categorization, document clustering and document classification. They differ in a number of aspects, such as: scientific tradition, methodology (including document pre-processing and indexing, test collections, characteristics of categories, evaluation methods) and application. However, all of them deal with the same problem and similarities between them exist; for example, selection of most relevant terms during document pre-processing is common to all the approaches, as is utilization of specific document characteristics. This leads one to assume that idea exchange and co-operation between the three communities would be beneficial.

The goal of the study is to examine whether simple bibliometric methods can be used to investigate to what degree the three communities utilize others' ideas, methods, and findings. Our main hypothesis is that there is hardly any exchange of ideas etc. To that purpose we studied direct links (do authors from one community cite authors from another) and indirect links using bibliographic coupling (Kessler, 1963). A freely available, offline tool for bibliometric analysis, Bibexcel, was used for the informetric analysis and map generation[2].

This paper is laid out as follows: brief descriptions of the three approaches are given in the next section, followed by a description of the methodology; results are discussed and conclusions are given in the last two sections.

Descriptions of the approaches

Text categorization is a machine-learning approach, in which also information retrieval methods are applied. It involves manually categorizing a number of documents to pre-defined

[1] The Swedish Agency for Innovation Systems and the Danish Ministry of Culture (grant no. A2004-06-028) has in part provided funding for this study. The authors wish to thank prof. dr. sc. Wolfgang Glänzel for his comments on the paper, given within a course organized by Nordic Research School in Library and Information Science (NORSLIS).

[2] Bibexcel is developed by prof. dr. sc. Olle Persson and may be dowloaded from http://www.umu.se/inforsk.

categories (which normally lack devices for the control of polysemy, synonymy and homonymy). By learning the characteristics of those documents the automated categorization of new documents takes place. Text categorization is known as supervised learning, since the process is "supervised" by learning categories' characteristics from manually categorized documents.

Document clustering is an information-retrieval approach. Unlike text categorization, it does not involve pre-defined categories or training documents and is thus called unsupervised. The clusters and, to a limited degree, relationships between clusters are derived automatically from the documents, and the documents are subsequently assigned to those clusters.

Document classification in this paper stands for a library science approach. It involves manually created controlled vocabulary (such as classification schemes, thesauri, or subject headings systems) into categories of which documents are classified. Controlled vocabularies have devices to control the problems of polysemy, synonymy and homonymy of natural language. They have been developed and used in libraries and in indexing and abstracting services, some since the end of the 19th century.

3 Methodology

Sample

The sample consists of 148 papers related to automated classification of Web-based text resources. The majority of papers are published after 1997. Out of these 63 papers are from the information retrieval (IR) community, 52 from machine learning (ML) and 33 from the library science (LS) community. The library science set of papers include two subgroups, one 'pure' library science subgroup, and the other with papers using either IR or ML approach, but also applying controlled vocabularies such as those used by the LS community.

The sample was collected from commercially and non-commercially available databases, mostly from ACM Digital Library, ISI Web of Science as well as Web sites of projects and personal Web sites. The databases were searched for documents on automated classification of text, using a variety of search terms. Not having any formal criteria, e.g. distinct channels of publication for each community, every paper had to be at least partially read in order to be assigned to the corresponding community. Additionally, due to overlaps in content, a number of papers were assigned to two or even three categories ('mixed' category in Table 1). For more than half of the papers, records with references had to be created from scratch or converted semi-automatically. The relatively small size of the sample is due to the fact that the number of LS papers is rather small (although there are many ML and IR papers).

Informetric methods used

Two main informetric methods were chosen for the study: direct and indirect links. *Direct links* were used to determine to what extent authors from one community cite authors from the other two communities. References of papers belonging to one community were searched for author names belonging to the other two communities. Every appearance of a name in the references was counted, which included different papers and even several instances where the searched author was cited as an editor of, e.g., conference proceedings. Only authors that were cited at least three times were examined. Authors, and not papers, were chosen because of the relatively small sample size. *Indirect links* were studied using bibliographic coupling between papers (Kessler, 1963). Bibliographic coupling was chosen because it shows the domain as it is interpreted by the researchers writing the new knowledge, and it is their own interpretation of their position in the scientific domain.

Results and discussion

Direct links between authors

Table 1 gives the number of times an author from one community was cited by one the other two communities, with percentages in the parentheses. They indicate that all communities cited each other. The IR and ML communities were mostly cited, and the LS community least cited.

Table 1. Number of citings between communities

	Authors cited by IR, excluding IR	**Authors cited by LS, excluding LS**	**Authors cited by ML, excluding ML**
IR	/	18 (24 %)	78 (42 %)
LS	7 (7 %)	/	29 (15 %)
ML	40 (41 %)	34 (45 %)	/
mixed	50 (52 %)	23 (31 %)	81 (43 %)

Qualitative analysis was used to determine the context in which other authors were cited. When the same paper from one community was cited by both other communities, it tended to be cited for similar reasons: either to provide an example of different classification methods and applications and compare with their own ones, or to refer to the same basic concepts of information retrieval and automatic text processing. Many of the authors citing other community's authors, also themselves belong to that community. The ML community uses IR methods and both tended to cite each other to a certain extent. LS authors cited by the other two communities did occur, but they were restricted to the 'non-pure' LS authors and papers. There was not one single case where 'pure' LS authors were cited by either of the two other communities, and vice versa: LS authors who cited the other two communities were either 'non-pure' or belonged to another community.

Indirect links between papers

Papers, as well as references, were identified by author and labelled with the author's name, her community's tag (IR, LS, ML) and publication year. Matrixes were produced in Bibexcel and imported into a multidimensional scaling (MDS) program for creation of two-dimensional maps. The stress of scaling was between 0.12 and 0.18, which indicated that the coupling was reasonably well reflected in the maps.

Only 110 of the 148 papers were bibliographically coupled. The majority of the pairs of papers with the largest number of mutually shared references belonged to ML community only, or both to ML and another one. Of LS-only papers that formed part of a coupled pair, all were 'non-pure' LS papers. Due to incomplete references, in several cases author name had to be 'replaced' by a made-up name (the same everywhere), in order for Bibexcel to work properly. Thus several pairs actually 'share' made-up authors. This could be corrected in the future by, for example, using different made-up names.

The MDS program has an upper limit on the number of papers that can be mapped. 62 papers were selected based on the following criteria: all mixed-category papers should be included; there should be an equal amount on papers in each category as far as possible; most frequently coupled papers should be included. Figure 2 shows the result of the mapping. Circle sizes indicate the total number of shared references for each paper, and lines between two papers

indicate that they are bibliographically coupled. The papers in the centre have many links with other papers. Those far down have lowest coupling frequencies. The same map is shown in Figure 3, but with the community tags only, and with lined groupings of the three communities. Papers belonging to several communities are left unmarked.

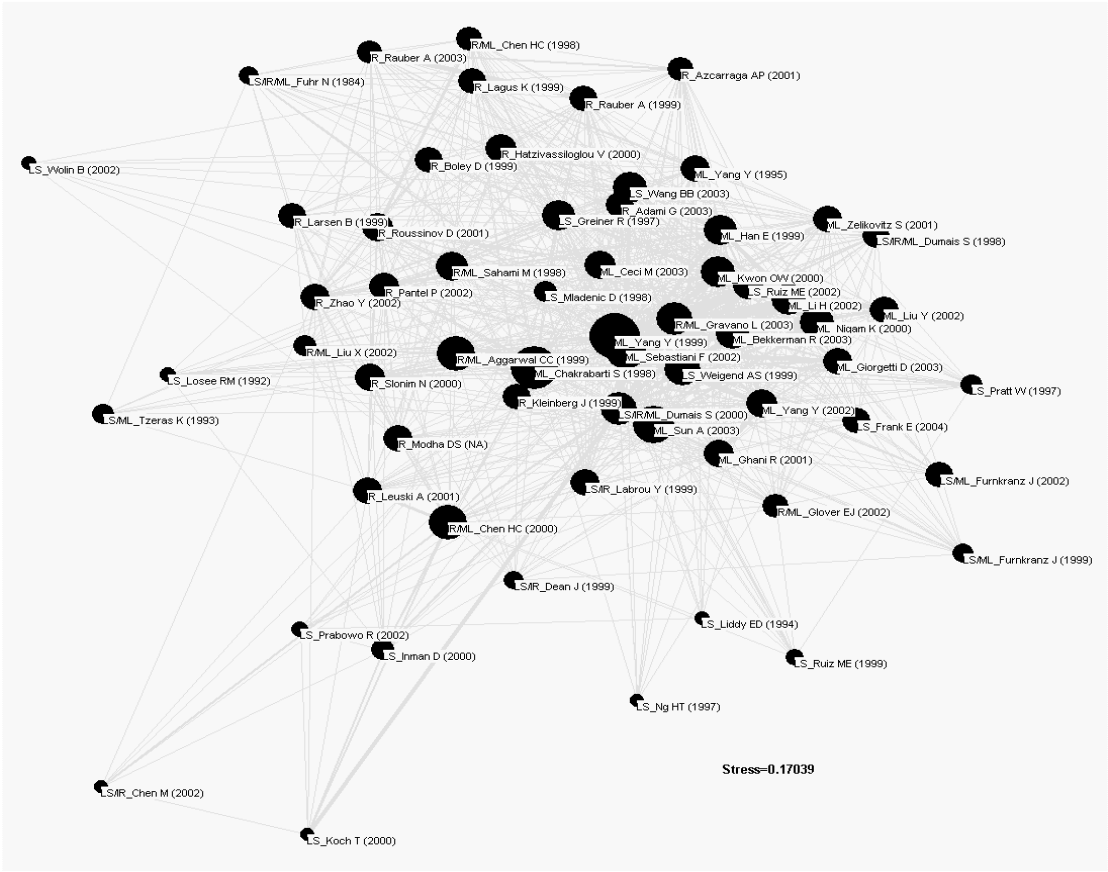

Figure 2: Bibliographic coupling map based on 62 selected papers, with circle sizes indicating the number of shared references

On both maps, ML papers are situated in the upper right corner and towards the centre, with IR papers continuing on their left, whereas LS papers are separated from the two of them and are positioned much lower, because they have lower coupling frequencies. LS papers are also much more scattered throughout the area, and connected with fewer lines to others because their coupling links are rarer. One can see that ML and IR are more closely related to each other than to the LS community. ML, and then the IR community, are most frequently coupled ones. Those LS papers with more links to IR and ML papers and with higher coupling frequencies belong to the 'non-pure' subgroup. The majority of mixed category papers are positioned close to either of the categories they were assigned to, indicating which group they belong to more.

Most clearly seen in Figure 3, the three communities form more or less distinct groupings. By further examining LS papers positioned between ML and IR areas, it was discovered that those were papers from the subgroup of LS coming from ML or IR but using a manually created vocabulary. This shows that even the group using controlled vocabularies couples with ML and/or IR, and not with other, 'pure' LS papers.

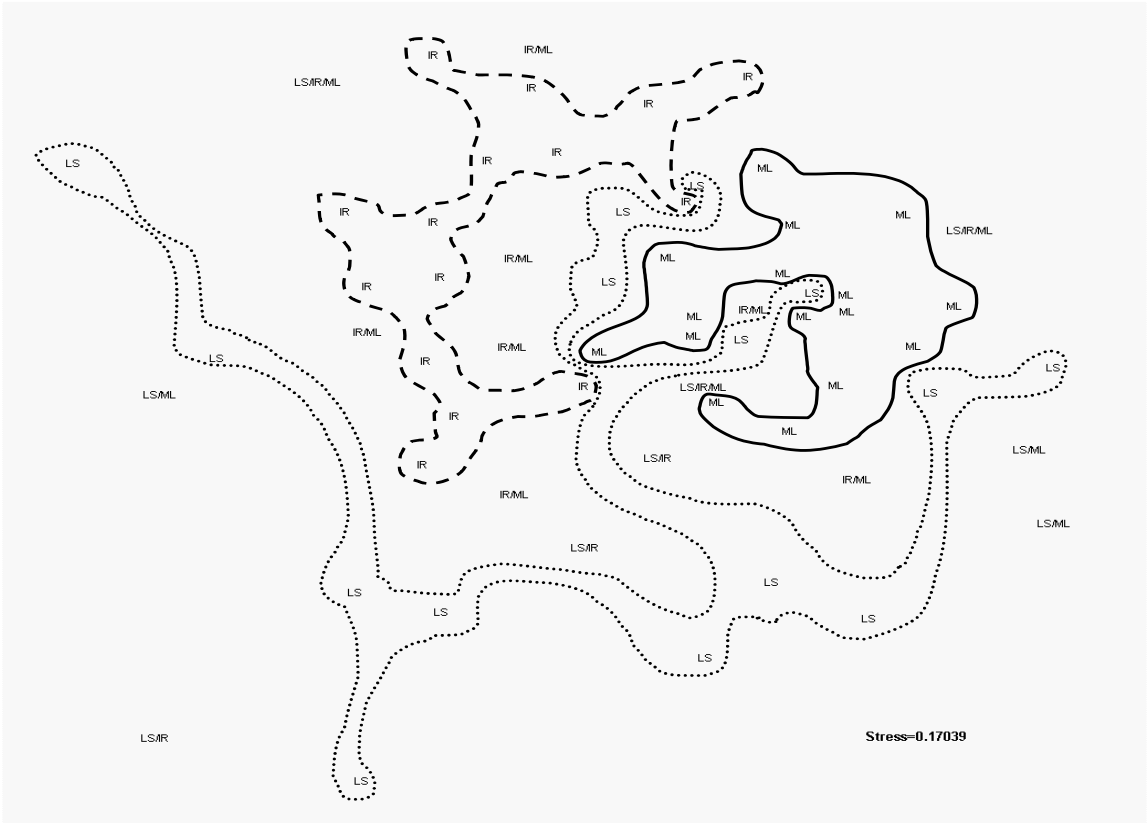

Figure 3: Bibliographic coupling map based on 62 selected papers with groupings emphasised

Conclusion

Using simple bibliometric methods on the sample of 148 papers, our hypothesis, that the three different communities researching automated classification do not communicate to a large extent, has been confirmed. Absence of ideas exchange was especially the case for the LS community, whereas the ML and IR community exchange ideas to a certain degree. The study of direct links showed that there was not a single case where 'pure' LS authors in the sample were cited by either of the two other communities. The situation was the same the other way around. ML and IR cited each other more but in many cases the authors citing another community's authors, themselves belonged to another community as well. Based on the bibliographic coupling analysis, one can see how the three communities form more or less distinct groupings. One could also see that ML and IR more closely related to each other than to the LS community. The LS and IR community were also most frequently coupled ones. It was discovered that those papers from the subgroup of LS coming from ML or IR but using a manually created vocabulary coupled with ML and/or IR, and not with other, 'pure' LS papers.

Further research would be based on a bigger sample and would deal, e.g., with the following questions: changing trends throughout different periods, and a more detailed analysis of why direct and indirect links are lacking between LS and the other two communities, in spite of appearance of ML and IR papers that employ controlled vocabularies.

Reference

Kessler, M. M. (1963). An experimental study of bibliographic coupling between technical papers. *IEEE transactions on information theory,* 9(1), 49-51.