Date : 25/05/2006

Challenges in automated classification using library classification schemes

Kwan Yi
School of Library and Information Science
University of Kentucky
USA

Meeting:	**97 Information Technology with Audiovisual and Multimedia and National Libraries (part 2)**
Simultaneous Interpretation:	No

WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL
20-24 August 2006, Seoul, Korea
http://www.ifla.org/IV/ifla72/index.htm

Abstract:
A major library classification scheme has long been standard classification framework for information sources in traditional library environment, and text classification (TC) becomes a popular and attractive tool of organizing digital information. This paper gives an overview of previous projects and studies on TC using major library classification schemes, and summarizes a discussion of TC research challenges.

1. INTRODUCTION

The enormous increase in the amount of digital information or resource available and the demand for retrieval tools to manage the information overload have lead to an interest in automatic classification task with the expectation of reducing human labor to a significant extent or even replacing in a limited portion. There have been a few research projects and some related studies on the feasibility of Library of Congress Classification (LCC) and Dewey Decimal Classification (DDC) as a classification framework for the automatic classification of digital information.

A major approach for organizing information is to classify collected information according to a pre-defined set of classes and to retrieve relevant information by browsing the list of classes used. This is a traditional way of classifying and locating library items based on library classification schemes. The old-fashioned was rekindled in digital environment with the popularity of subject directory and Web directory. However, a challenging issue of the approach is in the lack of having an authoritative classification schemes. The adoption of a library classification scheme appears to be a strong potential to fill the gap, supported by the practical popularity as a traditional classification role and the systematic and theoretical foundation. Considered the current issue, the new research agenda seems to be a promising research area for digital information organization and retrieval.

As a broad range of related and supportive studies and applications have flourished, such as richness of completed and on-going digital library projects, and successful application of machine learning approach to the Information Retrieval field, the situation has been mature enough to reach a point where it is appropriate to evaluate achieved progress in the agenda of automatic classification and to identify current challenges for establishing research agenda for the next years.

Section 2 describes some background on Text Classification. Section 3 describes the overview of recent studies and projects on TC using library classification schemes. Section 4 discusses current challenges in the automatic information organization and Section 5 makes conclusions.

2. BACKGROUND

2.1 Understanding of Text Classification

As a relatively new research field spun off from the Information Retrieval (IR) research, text classification (TC) is a task of classifying documents to a pre-defined set of classes without human assistance. The task is quite similar to a subtask of subject cataloguing in traditional libraries, but most distinctive in automatic classification, rather than being manually done by professionals. TC has become much more attractive than ever as the need of information organization tools to cope with a vast amount of digital information is more pressing.

TC is known as labeling documents with most relevant classes chosen from a group of candidate classes. The three primary components are involved in the process of TC. The first component is objects to be classified, which are textual documents. Let $D = \{d_1, d_2, \Lambda\ , d_n\}$ be a set of documents. The second component is target classes in

consideration. Let $C = \{c_1, c_2, \Lambda\ , c_m\}$ be a set of target classes. The third component is a mapping algorithm that acts as a classifier. A mapping algorithm can be specified as a function taking a document as an input and producing a binary decision whether the document falls into a given class. The function is depicted with $F : d_i \xrightarrow{c_j} \{0,1\}$, where $1 \le i \le n$ and $1 \le j \le m$. The output of 1 means that the document is interpreted to fall into the considering class, and it is interpreted not to fall into the class with the output of 0. Hence, the realization of a mapping function *F* and the quality determine the performance of the TC as the function serves to measure the relevance of a document given a class.

TC tasks can be divided into different types, according to the number of classes and the number of class labels. If there are only two classes to be considered, i.e., the value of *m* in the set *C* is equal to 2, such a TC is said to be a *binary* classification task. With more than two classes, i.e., the value of *m* is larger than 2, it is said to be a *multi-class* classification. Also, if each document is associated with only one class label, it is called a *binary-label* classification. In *multi-label* classification, there is at least one class label associated with each document.

2.2 Machine Learning Approach to TC Applications

A primary concern in the research of TC is how a *machine* (generally refer to a computer system but the term *machine* is used in convention) acquires necessary knowledge intended for correct classification. A most dominant approach to this issue is machine learning (ML). A general learning framework of ML is that a machine gains knowledge from previous experience. A ML framework may be described as a systematic process consisting of the four components (Kubat, Bratko, and Michalski 1999): *experiences*, *background knowledge*, *learning algorithm*, and *target knowledge*. *Experiences* as an input of the framework intend to convey what to learn for *target knowledge*, called *explicit* knowledge. *Background knowledge* as another input refers to the prior knowledge for the target class, called *implicit* knowledge. *Learning algorithm* as a black box of the ML framework represents a method of learning knowledge. *Target knowledge* is the realization of what a ML model learns from the combination of explicit and implicit knowledge. For example, when chess game is considered to be a learning task, a sequence of the chessboard positions changed in practice can be examples for the chess task, and general chess rules could be background knowledge. The ML approach is to some extent similar to the learning process by human beings. As humans may gain knowledge by reading documents, a machine in ML also acquires knowledge of a topic or class from documents that were pre-selected by domain specialists. Such a collection of documents provided is called a *training* set, whether it is explicit or implicit.

ML techniques have been applied to the classification of various types of documents: in-health documents (Larkey and Croft 1996), flora data (Cui, Heidorn, and Zhang 2002), legal documents (Thompson 2001), and Web documents (Chakrabarti, Dom, and Indyk 1998). Also, other types of category, other than subject or topic, were sought: document genres such as editorial, report, review, research paper, and homepage (Lee and Myaeng 2002), essay grading on various disciplines (Larkey 1998), filtering spam-mails (Hidalgo, López, and Sanz 2000).

2.3 Scope of Text Classification

The discussion of TC is in general bounded by the following:

- TC and *clustering* are similar but different mechanism in document classification. The concept of clustering is akin to TC, in grouping similar documents together. A clustering is defined as "*the group of documents which satisfy a set of common properties*" (Baeza-Yates and Ribeiro-Neto 1999). In clustering, any explicit set of classes are not taken into consideration. Instead, unknown common characters are sought. Therefore, In TC, the degree of similarity between a document and a target class is measured to see how relevant the document is to the class. In clustering, similar documents are not measured against target classes, but against other documents.

- Non-textual features (other clues than text) are beyond being considered.

- TC is content-based classification, is not based on metadata and structured information. TC and Web document classification are distinct in the use of metadata other than Web contents, such as hyperlinks and structured data, in Web classification.

3. OVERVIEW OF TEXT CLASSIFICATION PROJECTS AND STUDIES WITH LIBRARY CLASSIFICATION SYSTEMS

We will review a number of recent efforts in automated classification of digital documents using major library classification schemes.

One of the pioneer works on automated classification based on a library classification system can be found in (Larson 1992), where a set of MARC records was classified into the class Z (Bibliography and Library Science) of LCC, based on title and subject headings with 30,471 MARC records for training and 286 MARC records for testing. It was speculated that this work would help librarians to determine relevant classification numbers for unclassified items by providing a list of potential classification numbers based on subject headings and titles. The most recent work directly linked to Larson's work can be found in (Frank and Paynter 2004). Their work aims to assign LCC to metadata of Internet resources using LCC and Library of Congress Subject Headings (LCSH). The classifier is trained using 800,000 library catalog records and tested on an independent set of 50,000 records. The classification accuracy of this classifier is reported as a wide range from 55% to 80%.

In the following sections, a number of TC projects where traditional library classification schemes were adopted as the basis for a classification system for digital documents will be reviewed.

3.1 Pharos

Pharos is an information architecture prototype accommodating heterogeneous sources in content and format, derived from the Alexandria Digital Library project (Dolin, Agrawal, and El Abbadi 1999). As an initial prototype of the Pharos

architecture, an automatic classification system based on the LCC was implemented for the purpose of creating the profiles of heterogeneous digital information. In this project, the Latent Semantic Indexing technique was used for automatically classifying newsgroups and cataloguing records within the LCC. As a training data set, 1.5 million catalogue records from the University of California Santa Barbara library were used, and title, subject headings, and LCC fields from the records were extracted. For a specific holding, title and subject heading data are viewed as a description for a specific category denoted by a LCC number. Such a relationship between a LCC number and its descriptors forms training data for the classification system. 7214 MARC records from the 21 major classes of LCC were classified, and the experimental results yielded an average median of 13.0 ± 3.9 and an average mean of 76 ± 19 for about 4,200 LC classes. In another experiment with articles from 2,500 Usenet newsgroups, the classification accuracy for the experiment is not reported since articles that were not pre-classified were involved.

3.2 Scorpion

Scorpion was a research project led by the Online Computer Library Center (OCLC) from 1996 to 1999 with an aim at developing an automated method of identifying the DDC of digital documents (Shafer 2001). For its automatic classification, Scorpion used a clustering method. Given an input document, Scorpion measures similarities between the input document and the pre-defined clusters (corresponding to DDC classes) and considers the nearest cluster as the most probable place for the input document. A term counting is used as a measurement of similarity. For the evaluation, a collection of bibliographic records for Internet resources in which DDC classes were human-assigned was used. Unfortunately, however, detailed experimental results were not unveiled, presumably because a comparison could not be properly done because the human-assignment of DDC classes was based solely on phrases describing Internet resources. Their conclusions confirmed that automatic classification cannot replace manual classification, but that it can provide a cost effective solution to support human catalogers.

3.3 DESIRE (Koch and Ardö 2000)

The DESIRE project, started in 1996, is a large-scale international project funded by the European Union for the purpose of building a subject gateway for engineering-subject resources. In an experiment, Web documents were automatically classified into the Engineering Information (EI) classification, which was relied on simple term matching. In the system's evaluation with approximately 1000 Web pages, the automatic classification's accuracy was compared against the classification staffs' decisions. Overall, the fact that about 60% of the automatic classifications were correctly or more finely matched to the human decisions was reported. With the collaboration of OCLC, the same engineering data was classified under DDC. Particularly, some LCSH were added in addition to the full-text. An evaluation for the DDC-based classification was not reported.

3.4 Wolverhampton Web Library (Jenkins et al.)

Wolverhampton Web Library (WWLib)[1] is a Web search engine project for UK-based documents, where DDC is used to organize the collected documents. An interesting feature of the experimental WWLib is to treat a Web page as an item in a library and to prepare cataloging records describing information including the title, Universal Resource Locator (URL), DDC category, and description to the collected Web pages. In general, Web search engines present results in the order of relevance to users' requests, whereas the WWLib provides the relevant Web pages in terms of DDC category. The Classifier component performs the process of classifying Web documents automatically, which relies on simple word matching. The classifier in the WWLib compares a stream of words extracted from documents and the description of DDC categories (Wallis & Burden, 1995). The words occurring in Web documents are *weighted* according to the tags used for them, and a stemming technique is applied. Also, to take advantage of the hierarchical structure of DDC, a method for the relevance of a document to both a class and its upper class is taken into consideration. In the later version (WWLib), a more rich set of description for DDC classes including synonyms is considered. A formal experiment for the measurement of the system's performance seems not to be undertaken. Instead, an informal testing result was reported with the randomly selected 17 URLs (WWLib); where 13 cases out of 17 were simply reported to be relevant without divulging more detailed procedures such as evaluation methods and data selection.

3.5 Summary

From the standpoint of application, we found that most research projects described above conducted automatic classification of cataloging data and Web pages, not applied to full-text documents. From the viewpoint of the classification scheme, either LCC or DDC, both of which are most popularly used library classifications in North America, has been used in classification applications. However, the rationale for choosing a library classification is not clearly explained. According to the descriptions written in articles, the decision for the choice of a library classification scheme seems to be based upon personal preference, rather than the tasks at hand, or the availability of data.

4. CHALLENGES IN TEXT CLASSIFICATION FOR LIBRARY CLASSIFICATION SYSTEMS

The discussion of TC challenges is broken down by the component of TC process.

4.1 Classification Schemes

Library classification schemes were originally developed to organize primarily printed materials such as books and serials, and this has been primarily used in traditional library settings for over a century. Recently, the use of the schemes is further extended into the online environment for organizing digital information where the potential role of library classification schemes has been explored as tools to

[1] http://www.scit.wlv.ac.uk/wwlib

organize, browse, and access information. A number of universal library classification schemes have been used in various projects and studies (Koch and Day 1997), such as the LCC, DDC, National Library of Medicine (NLM), and Universal Decimal Classification (UDC).

4.1.1 Class Coverage and Characteristics

The first challenge comes with the size and volatility of classes. There are approximately 100,000 different classes in LCC and the class number of DDC is not far from it. Thus, preparing training data for each class and constructing a TC system corresponding to each class do not seem to be logistically possible. Moreover, classification schemes are not static all the time, but are being re-examined including revisions of existing classes. Also, all the classes specified in classification systems would not seem to be used at all.
The second challenge is in dissimilarity of classification schemes. Universal library classification schemes are in common that subject is the primary characteristic for classes. However, they are quite different in structural nature and notation system adopted.

To cope with the issues, the followings may be taken into considerations.

- Set the limits of class level according to the broad topic of TC application
    - Different TC applications are interested in different level or set of classes. A history application may be interested in the class of history, whereas the use of full range of classes is more attractive to an application like web directory.
- Implement the characteristics of classification structure
    - In classification schemes, the relationships among classes are reflected in their hierarchical structure. That is, the DDC is a hierarchical classification in that a class in a level indicates a more general discipline or subject than a class in its subordinate level. The nature of the LCC hierarchy is similar to that of the DDC. A set of main classes on the top of the hierarchy represents a list of disciplines, and each of them is divided into subclasses for more specific disciplines, except the E and F (history in America), and Z (bibliography and library science) classes. Then, further subdivisions are generally made by topic, place, time, and form. In this way, training data for lower levels of class can also be used for higher levels.

4.2 Source of Knowledge

To make it simple, the target knowledge acquired by machine is directly derived from training data set as input, and the resultant knowledge by TC process may be inherently affected by that of training data. Thus, TC systems target for same knowledge ends in different knowledge acquired when different training data set are used.
In general, the acquisition of training data is known as a difficult, labour-intensive, and cost expensive process, and may be infeasible in some cases. Training data consist of *experiences* and background knowledge. Background knowledge is a

general data set applicable to much broader subject, whereas experience is held for a task-oriented specific subject or class.

4.2.1 More explicit training data for testbed

ML-based TC research depends on training data. Today there are a few of standard training data used as benchmark for the research. Developing more training data in diverse subjects may spur much productive research and lead to major improvement in TC.

4.2.2 Developing background knowledge

Often confronted with the scarcity of training data and difficulty to access them, research aimed at developing tools for generating background knowledge is significant. Controlled vocabularies and thesauri are collections of authorized terms in subject areas and also represent some types of term relationships. Potential of using their definitions and relationships for background knowledge look to be prominent. Also, Implementation of integrating and interlinking various knowledge organization tools and sources may significantly improve.

4.2.3 Automatic data generation

Alternative to manual data collection would be substantially useful since data generation is the most costly process in TC. The information environment for this is mature: availability of a vast volume of digital information; and a broad range of information processing tools and techniques developed in information retrieval field. Digital resources pre-classified by professionals can be often found in some information systems, such as Web directories and online databases. Use of highly pre-controlled documents may be an option for new data collection.
Once a TC system sets up, it takes un-classified documents as input and produce classified documents as output. Another possibility is in the re-use of the classified documents as training data set.

4.2.4 Integration of multiple sources

Models and tools for incorporating multiple sources of evidence such as background knowledge and explicit knowledge through varied routes may be significantly in synthesizing clues for relevant classes.

4.3 Classification Techniques/Models

A growing number of researchers from various fields of study, primarily in computer and information science have been interested in the development of automated text-based document classification tools and methods. A broad range of inductive learning algorithms and techniques, such as Support Vector Machines, Baysian Belief Network, Decision Trees, and Artificial Neural Networks, have been proposed and tested, (Yiming 1994; Joachims 1998; Lewis and Ringuette 1994; Mitchell 1997). TC research has heavily focused on the development of effective techniques, methods, and learning algorithms (Sebastiani 2002).

Different types of classification models support different representations of knowledge, and adopt different learning methods. In neural network algorithms, knowledge is represented as a graph consisting of nodes and edges, and, in rule induction, condition-action rules are used. In other methods, functions, logic programs and rule sets, finite-state machines, grammars, and problem solving systems have been adopted to represent knowledge.

4.3.1 Semantic Indexing Techniques and Classification Models

An ultimate goal in text classification may be the full understanding of meanings of textual documents. This challenging issue has long been tackled by researchers primarily from computational linguistics and natural language understanding, since the beginning of automatic document processing dated back to 1950s.
TC, as a subfield of Information Retrieval (IR) research, has adopted various IR techniques and models developed. Recently information researchers in information processing turned to artificial intelligence-based learning methods and tools from neural networks, symbolic learning, and genetic algorithms. Probabilistic techniques and methods for TC have been more attracted in the past decade.
Current models of text deal with relatively simple aspects of language (words, phrases, names), and models for indexing terms rely on simple counting mechanism such as frequency and co-occurrence. Such models do not capture aspects of semantic structure and relationships that may be more important for characterizing TC classes in subject, topic, etc. Typical examples for the problems with non-semantic methods can be found in word synonym (multiple words with a same meaning) and polysemy (a word with multiple meanings).
A number of efforts to resolve the issue in part or as a whole has been attempted in IR fields. A promising direction seems to be to explore models incorporating authority controlled terms and relationships among them.

5. CONCLUSION

During the last century, the role of library classification schemes have been expanded as tools for locating library holdings on shelves, for browsing them through Online Public Access Catalog (OPAC), and now for organizing and accessing digital resources in networked environments. The adoption of traditional classification schemes to digital environment is attractive, promising, and potential for the following reasons: (1) Major library classification schemes have been most popularly used information organization frameworks; (2) A rich set of organization tools supporting and in association with library classification schemes, such as controlled vocabularies and subject headings, has been developed and available; (3) Bibliographic descriptions of information sources such as cataloging records contains the association between information sources and bibliographic tools used.

References

Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval* New York, NY ACM Press.

Chakrabarti, S., B. E. Dom, and P. Indyk. 1998. Enhanced hypertext categorization using hyperlinks. Paper read at The ACM SIGMOD, at Seattle, WA.

Cui, Hong, P. B. Heidorn, and Hong Zhang. 2002. An approach to automatic classification of text for information retrieval. Paper read at The 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, at Portland, Oregon.
Dolin, R., D. Agrawal, and A. El Abbadi. 1999. Scalable collection summarization and selection, at Berkley, CA, USA.
Frank, E., and G. W. Paynter. 2004. Predicting Library of Congress classifications from Library of Congress subject headings. *Journal of the American Society for Information Science and Technology* 55 (3):214-227.
Hidalgo, J. G., M. Maña López, and E. Puertas Sanz. 2000. Combining Text and Heuristics for Cost-Sensitive Spam Filtering. Paper read at The Fourth Computational Natural Language Learning Workshop, at Lisbon, Portugal.
Jenkins, Charlotte, Mike Jackson, Peter Burden, and Jon Wallis. 2006. *Automatic classification of Web resources using Java and Dewey Decimal Classification* [cited May 12 2006]. Available from http://www7.scu.edu.au/1846/com1846.htm.
Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Paper read at The 10th European Conference on Machine Learning, April 21-24, 1998, at Chemnitz, Germany.
Koch, Traugott, and Anders Ardö. 2006. *Automatic classification: DESIRE II D3.6a, Overview of results* 2000 [cited May 12 2006]. Available from http://www.lub.lu.se/desire/DESIRE36a-overview.html.
Koch, Traugott, and Michael Day. *The role of classification schemes in Internet resource description and discovery* 1997 [cited. Available from http://www.ub2.lu.se/desire/radar/reports/D3.2.3/
Kubat, Miroslav, Ivan Bratko, and Ryszard S. Michalski. 1999. A Review of Machine Learning Methods. In *Machine Learning and Data Mining: Methods and Applications*, edited by R. S. Michalski, I. Bratko and M. Kubat. Chichester, England: John Wiley & Sons.
Larkey, L. S. 1998. Automatic essay grading using text categorization techniques. Paper read at The 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, at Melbourne, Australia.
Larkey, L. S., and W. B. Croft. 1996. Combining classifiers in text categorization. Paper read at The 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 18-22, 1996, at Zurich, Switzerland.
Larson, R. R. 1992. Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science* 43 (2):130-148.
Lee, Yong-Bae, and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features, at Tampere, Finland.
Lewis, D. D., and M. Ringuette. 1994. A Comparison of Two Learning Algorithms for Text Categorization. Paper read at The 3rd Annual Symposium on Document Analysis and Information Retrieval, at Las Vegas, NV.
Mitchell, Tom M. 1997. *Machine Learning*. Boston, MA: McGraw-Hill.
Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34 (1):1-47.
Shafer, Keith E. 2001. Evaluating Scorpion results. OCLC research project using DDC for automatic subject assignment. *Journal of Library Administration* 34 (3/4):237-44.

Thompson, Paul. 2001. Automatic categorization of case law. Paper read at The 8th International Conference on Artificial Intelligence and Law, May 2001, at St. Louis, Missouri.
Yiming, Yang. 1994. Expert Network: effective and efficient learning from human decisions in text categorization and retrieval, at Dublin, Ireland.