Leveraging the Legacy of Conventional Libraries for Organizing Digital Libraries

Arash Joorabchi and Abdulhussain E. Mahdi

Department of Electronic and Computer Engineering, University of Limerick, Ireland
{Arash.Joorabchi,Hussain.Mahdi}@ul.ie

Abstract. With the significant growth in the number of available electronic documents on the Internet, intranets, and digital libraries, the need for developing effective methods and systems to index and organize E-documents is felt more than ever. In this paper we introduce a new method for automatic text classification for categorizing E-documents by utilizing classification metadata of books, journals and other library holdings, that already exists in online catalogues of libraries. The method is based on identifying all references cited in a given document and, using the classification metadata of these references as catalogued in a physical library, devising an appropriate class for the document itself according to a standard library classification scheme with the help of a weighting mechanism. We have demonstrated the application of the proposed method and assessed its performance by developing a prototype classification system for classifying electronic syllabus documents archived in the Irish National Syllabus Repository according to the well-known Dewey Decimal Classification (DDC) scheme.

Keywords: Digital library organization, text classification, collective classification, library classification schemes, bibliography.

1 Introduction

Similar to conventional libraries, large-scale digital libraries contain hundreds of thousands of items and therefore require advanced querying and information retrieval techniques to facilitate easy and effective search systems. In order to provide highly precise search results, such search systems need to go beyond the traditional keyword-based search techniques which usually yield a large volume of indiscriminant search results irrespective of their content. Classification of materials in a digital library based on a pre-defined scheme could improve the accuracy of information retrieval significantly and allows users to browse the collection by subject [1]. However, manual classification of documents is a tedious and time-consuming task which requires an expert cataloguer in each knowledge domain represented in the collection, and therefore deemed impractical in many cases. Motivated by the ever-increasing number of E-documents and the high cost of manual classification, Automatic Text Classification/Categorization (ATC) - the automatic assignment of natural language text documents to one or more predefined categories or classes according to their contents - has become one of the key methods to enhance the information retrieval and knowledge management of large digital textual collections.

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 3–14, 2009.

Until the late ’80s, the use of rule-based methods was the dominant approach to ATC. Rule-based classifiers are built by knowledge engineers who inspect a corpus of labeled sample documents and define a set of rules which are used for identifying the class of unlabelled documents. Since the early ’90s, with the advances in the field of Machine Learning (ML) and the emergence of relatively inexpensive high performance computing platforms, ML-based approaches have become widely associated with modern ATC systems. A comprehensive review of the utilization of ML algorithms in ATC, including the widely used Bayesian Model, k-Nearest Neighbor, and Support Vector Machines, is given in [2]. In general, an ML-based classification algorithm uses a corpus of manually classified documents to train a classification function which is then used to predict the classes of unlabelled documents. Applications of such algorithms include spam filtering, cataloguing news and journal articles, and classification of web pages, just to name a few. However, although a considerable success has been achieved in above listed applications, the prediction accuracy of ML-based ATC systems is influenced by a number of factors. For example, it is commonly observed that as the number of classes in the classification scheme increases, the prediction accuracies of the ML algorithms decreases. It is also well-recognized that using sufficient number of manually classified documents for training influences the prediction performance of ML-based ATC systems considerably. However, in many cases, there is little or no training data available. Hence, over the last few years, research efforts of the machine learning community has been directed towards developing new probability and statistical based ML algorithms that can enhance the performance of the ML-based ATC systems in terms of prediction accuracy and speed, as well as reduce the number of manually labeled documents required to train the classifier.

However, as Golub [3] and Yi [4] discuss, there exits a less investigated approach to ATC that is attributed to the library science community. This approach focuses less on algorithms and more on leveraging comprehensive controlled vocabularies, such as library classification schemes and thesauri that are conventionally used for manual classification of physical library holdings. One of the main applications of this approach to ATC is in the automatic classification of digital library holdings, where using physical library classification schemes is a natural and usually most suitable choice. Another application is in classifying web pages, where due to their subject diversity, proper and accurate labeling requires a comprehensive classification scheme that covers a wide range of disciplines. In such applications using library classification schemes can provide fine-grained classes that cover almost all categories and branches of human knowledge.

A library classification system is a coding system for organizing library materials according to their subjects with the aim of simplifying subject browsing. Library classification systems are used by professional library cataloguers to classify books and other materials (e.g., serials, audiovisual materials, computer files, maps, manuscripts, realia) in libraries. The two most widely used classification systems in libraries around the world today are the Dewey Decimal Classification (DDC) [5] and the Library of Congress Classification (LCC) [6]. Since their introduction in late 18th century, these two systems have undergone numerous revisions and updates.

In general, all ATC systems that have been developed using above library science approach can be categorized into two main categories:

1. String matching-based systems: these systems do not use ML algorithms to perform the classification task. Instead, they use a method which involves string-to-string matching between words in a term list extracted from library thesauri and classification schemes, and words in the text to be classified. Here, the unlabelled incoming document can be thought of as a search query against the library classification schemes and thesauri, and the result of this search includes the class(es) of the unlabelled document. One of the well-known examples of such systems is the Scorpion project [7] by the Online Computer Library Centre (OCLC). Scorpion is an ATC system for classifying E-documents according to the DDC scheme. It uses a clustering method based on term frequency to find the most relevant classes to the document to be classified. A similar experiment was conducted by Larson [8] in early 90's, who built normalized clusters for 8,435 classes in the LCC scheme from manually classified records of 30,471 library holdings and experimented with a variety of term representation and matching methods. For more examples of these systems see [9, 10].
2. Machine learning-based systems: these systems use ML-based algorithms to classify E-documents according to library classification schemes such as DDC and LCC. They represent a relatively new and unexplored trend which aims to combine the power of ML-based classification algorithms with the enormous intellectual effort that has already been put into developing library classification systems over the last century. Chung and Noh [11] built a specialized web directory for the field of economics by classifying web pages into 757 sub-categories of economics category in DDC scheme using k-NN algorithm. Pong et al. [12] developed an ATC system for classifying web pages and digital library holdings based on the LCC scheme. They used both k-NN and Naive Bayes algorithms and compared the results. Frank and Paynter [13] used the linear SVM algorithm to classify over 20,000 scholarly Internet resources based on the LCC scheme. In [14], the authors used both Naïve Bayes and SVM algorithms to classify a dataset of news articles according to the DDC scheme.

In this paper, we propose a new category of ATC systems within the framework of the "library science" approach, which we call Bibliography Based ATC (BB-ATC). The proposed BB-ATC system uses a novel, in-house developed method for automatic classification of E-documents which is solely based on the bibliographic metadata of the references cited in a given document and have been already classified and catalogued in a physical library.

The rest of the paper is organized as follows: Section 2 describes the proposed BB-ATC method. Section 3 describes a prototype ATC system which has been developed based on the proposed method in order to demonstrate the viability of proposed method and evaluate its performance. Section 4 describes details of above evaluation and its experimental results. This is followed by Section 5 which analyses presented results and discusses some of the main factors affecting our classifier's performance. Section 6 provides a conclusion and discusses future work.

2 Outline of Proposed BB-ATC Method

A considerable amount of E-documents have some form of linkage to other documents. For example, it is a common practice in scientific articles to cite other articles and books. It is also common practice for printed books to reference other books, documented law cases to refer to other cases, patents to refer to other patents, and webpages to have links to other webpages. Exploring the potential of leveraging these networks of references/links for ATC opens a new route for investigating the development of ATC systems, which can be linked to the field of collective classification [15]. Our proposed BB-ATC method falls into this route, and aims to develop a new trend of effective and practical ATC systems based on leveraging:

- The intellectual work that has been put into developing and maintaining resources and systems that are used for classifying and organizing the vast amount of materials in physical libraries; and
- The intellectual effort of expert cataloguers who have used above resources and systems to manually classify millions of books and other holdings in libraries around the world over the last century.

With the assumption that materials, such as books and journals, cited/referenced in a document belong to the same or closely relevant classification category(ies) as that of a citing document, we can classify the citing document based on the class(es) of its references/links as identified in one or more existing physical library catalogues. The proposed BB-ATC method is based on automating this process using three steps:

1. Identifying and extracting references in a given document;
2. Searching one or more catalogues of exiting physical libraries for the extracted references in order to retrieve their classification metadata;
3. Allocating a class(es) to the document based on retrieved classification category(ies) of the references with the help of a weighting mechanism.

It should be noted here that this method is applicable to any E-document that cites/references one or more items which have been catalogued in the library catalogues searched by the system. Examples of such E-documents include books, journal and conference papers, learning and teaching materials (such as syllabi and lecture notes), theses and dissertations to name a few.

In order to make a viable ATC system, the proposed method needs to adopt a specific standard library classification scheme. For the purpose of this work, both the Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) schemes were considered as candidate classification schemes, due to their wide use and subject coverage. We adopted the DDC in our BB-ATC method for two main reasons:

- The DDC scheme is used for classifying library holdings in about 80% of libraries around the world and, therefore, the number of existing items that are classified according to the DDC is much greater than those classified according to the LCC. This makes the DDC scheme a better choice for our method which is based on utilizing the classification metadata of items that have been manually classified according to a library classification scheme.

- The DDC has a fully hierarchical structure while the LCC is not fully hierarchical and usually leans toward alphabetic or geographic sub-arrangements. The hierarchical relationships among topics in the DDC are clearly reflected in the classification numbers which can be continuously subdivided. The hierarchical feature of the DDC allows the development of effective GUI interfaces that enable users to easily browse and navigate the scheme to find the categories that they are interested in without requiring prior knowledge of the classification scheme or its notational representation.

3 System Implementation and Functionality

In order to demonstrate the viability and performance of the proposed BB-ATC method, we have developed a prototype ATC system for classifying electronic syllabus documents for the Irish National Syllabus Repository [16] according to the DDC scheme. Figure 1 shows an overview of the system. As illustrated, the system is effectively a metadata generator comprising a Pre-processing unit, an Information Extractor, a Catalogue-search unit, and a Classifier.

The system has been designed such that it can handle syllabus documents of different formats, such as PDF, MS-Word and HTML. To facilitate this, a pre-processing unit with the help of Open Office Suite [17] and Xpdf [18] extracts the textual content of the arriving documents and passes the results to the rule-based Information Extractor (IE). The IE uses the JAPE transducer from GATE toolkit [19] to identify and extract the ISBN number of each book that has been referenced in the document. After extracting the unique identifiers of all referenced items, the Catalogue-Search unit of the system uses the Z39.50 protocol [20] to create a search query for each reference based on its unique identifier. The search query is then submitted to the OPAC (Online Public Access Catalogue) search engines of the Library of Congress (LOC) and the British Library (BL) to search for matching records. The returned search results contain the records of matching catalogued items in MARC21 format [21]. Each record holds bibliographic and classification metadata about a book including its DDC classification number and the Library of Congress Subject Headings (LCSHs) assigned to it. The Catalogue-Search unit iterates through the retrieved records and extracts the classification numbers and LCSHs of all the referenced materials and passes them to the Classifier.

The task of the Classifier is to use the retrieved classification metadata of references in the document to classify it into one or more classes from the DDC using a simple weighting mechanism. The weighting mechanism works as follows: if the document to be classified contains only one reference belonging to a single class, then the document is assigned to that single class. The same applied if the document contains more than one reference, but all the references belong to the same class. However, if the references used in the document belong to more than one class, then the weight of each class is equal to the number of references in the document which belong to that class. In this case the document is labelled with all the relevant classes and their corresponding weights. This weighting mechanism enriches the classification results by providing not only a list of classes relevant to the subject(s) of the document but also a relevance measure for these relations. In addition to labelling the document with one or more classes from the DDC, the Classifier also assigns a weighted list of the Library of Congress Subject Headings (LCSHs) to the document. This list contains a set of all the LCSHs assigned to the references in the document

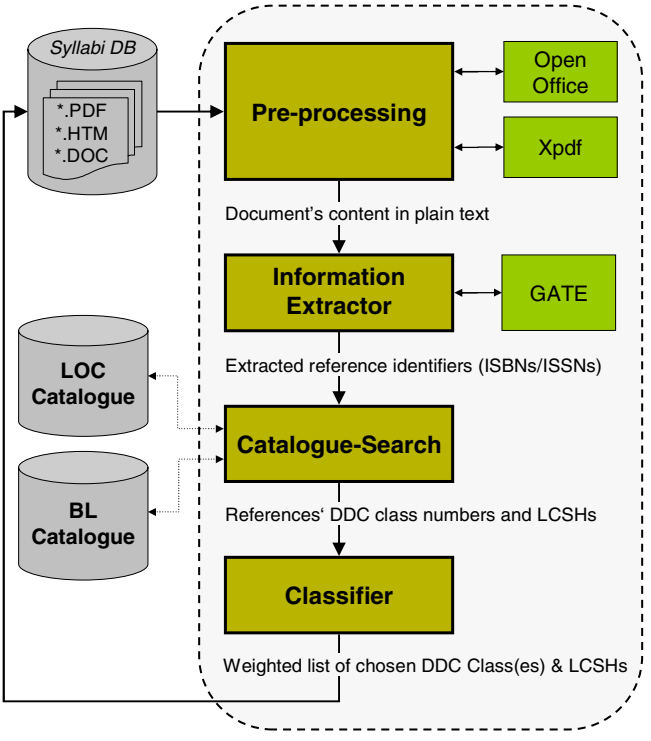

Fig. 1. Overview of the prototype ATC system

and their corresponding weights. The weight of each subject heading is equal to the number of times it has been assigned to the references. The LCSHs attempt to evaluate the subject content of items, whereas the DDC and the LCC rather broadly categorise the item in a subject hierarchy.

4 System Evaluation and Experimental Results

We used one hundred computer science related syllabus documents to evaluate the performance of our prototype ATC system. The documents have already been categorized by the expert cataloguers of the Irish National Syllabus Repository. We used the standard measures of Precision, Recall and their harmonic mean, F1, to evaluate the prediction performance of the system. Precision is the number of times that a class label has been correctly assigned to the test documents divided by the total number of times that class has been correctly or incorrectly assigned to the documents. Recall is the number of times a class label has been correctly assigned to the test documents divided by the number of times that class should have been correctly assigned to the test documents. Accordingly:

$$\text{Recall} = \frac{\#\ \text{Correctly assigned class labels}}{\#\ \text{Total possible correct}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\#\ \text{correctly assigned class labels}}{\#\ \text{Total assigned}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

where the Recall, Precision and F1 are computed in terms of the labels TP (True Positive), FP (False Positive), and FN (False Negative) to evaluate the validity of each class label assigned to a document, such that:

- TPi: refers to the case when both the classifier and human cataloguer agree on assigning class label i to document j;
- FPi: refers to the case when the classifier has mistakenly (as judged by a human cataloguer) assigned class label i to document j;
- FNi: refers to the case when the classifier has failed (as judged by a human cataloguer) to assign a correct class label i to document j.

In order to obtain a true objective evaluation of the classification performance, we applied the micro-average to above target performance measures (i.e., precision, recall, and F1) over all categories. Table 1 summarizes the achieved results.

Table 1. Micro-averaged classification results

TP	FP	FN	Precision	Recall	F1
210	19	26	0.917	0.889	0.902

Table 2 shows the classification results of a sample test syllabus document. The full and individually analyzed classification results of the one hundred test syllabus documents may be viewed online on our webpage[1]. We encourage readers to refer to this document as it provides detailed insight into the performance of the proposed BB-ATC method.

As a common practice in developing a new system, it is always desired to compare its performance to other existing similar systems. However, we would like to note here that it is not possible to conduct a true objective comparison between the performance of our method and other reported ATC systems that use either the string-matching or ML based approaches. This is due to the following:

- Unlike our system which classifies E-documents based on the full DDC scheme, other reported ATC systems, due to their limitations, either adopt only one of the main classes of DDC or LCC as their classification scheme or use an abridged version of DDC or LCC by limiting the depth of hierarchy to second or third level.
- In quite a few reported cases, the performance of the system was evaluated using different measures other than the standard performance measures of precision, recall, and F1 used in our case.

[1] http://www.csn.ul.ie/~arash/PDFs/1.pdf

Table 2. A sample syllabus classification results

Syllabus Title: Computer Graphics Programming		
DDC No. [***Weight***]	006.696 [1]	006.6 Computer graphics 006.69 Special topics in computer graphics **006.696 Digital video**
	006.66 [2]	006.6 Computer graphics **006.66 Programming**
LCSH [***Weight***]	• Computer graphics [3] • OpenGL [2]	• Computer Animation [1] • Three-dimensional display systems [1]

Despite above, it is possible to provide a relative comparison between the performance of our system and those of similar reported systems. For example, Pong and co-workers [12] used both the Naive Bayes and k-NN ML algorithms to classify 254 test documents based on a refined version of LCC scheme which consisted of only 67 categories. They reported the values of 0.802, 0.825, and 0.781 as the best figures for micro-averaged F1, recall, and precision, respectively, achieved by their system. Also, Chung and Noh [11] reported the development of a specialized economics web directory by classifying a collection of webpages, belonging to the field of economics, into 575 subclasses of the DDC main class of economics. Their unsupervised string-matching based classifier achieved an average precision of 0.77 and their supervised ML based classifier achieved an average precision and recall of 0.963 and 0.901, respectively. In [16] the authors used the naïve Bayes classification algorithm to automatically classify 200 syllabus documents according to the International Standard Classification of Education scheme [22]. The performance of the classifier was measured using one hundred undergraduate syllabus documents and the same number of postgraduate syllabus documents taken from the Irish National Syllabus Repository, achieving micro-average values of 0.75 and 0.60 for precision for each of the above document groups, respectively.

5 Discussion of Results

We examined each individual syllabus document in the test collection in conjunction to its predicted classification categories in more depth in order to obtain an insight into the behavior of the method and underlying factors that affect its performance. We first read the full content of each document and then examined the books that are referenced in the document and the DDC classes that these books have been assigned to by expert cataloguers in the Library of Congress and the British Library. In this section we summarize the findings of this examination.

- The majority of the DDC classes and LCSHs assigned to the documents are quite relevant to the contents of the documents and provide detailed and semantically rich information about the core subjects of the syllabi. For example, in case of the

sample syllabus document in Table 2, titled "computer graphics programming", the two classes assigned to the document, *digital video* (descended from *Special topics in computer graphics*) and *programming* (descended from *computer graphics*), objectively cover the two main subjects of the syllabus with the weight assigned to each class logically justifying this classification.

- The number of books referenced in test documents ranges from 1 to 10. In total, the one hundred test documents reference 365 books. 305 of these referenced books were catalogued and classified based on the DDC in either the LOC or the BL catalogue. We did not encounter a case in which none of the referenced books were catalogued. In 38 cases one or more of the referenced books were not catalogued in either the LOC or the BL. The results show that the existence of one or more (up to the majority) un-catalogued references in a document does not always lead to a misclassification of that document. However, as the number of catalogued references in the document increases recall of our system improves and the weights assigned to the classes and LCSHs become more accurate.
- The Dewey classes and the LCSHs are independent from each other. Therefore, if the class label of a referenced book does not match any of the main subjects of the document, i.e. an FP case, this does not mean that the LCSHs assigned to that reference are irrelevant too. In fact, in a substantial number of examined cases, many of the assigned LCSHs were quite relevant despite the fact that the classes of the referenced books were irrelevant to the subject of the documents. This is due to the fact that LCSHs are not bound to DDC classes and they provide subject labels that are not represented in the classification scheme exclusively or are not the main subject of the item being classified. For example, in case of syllabus document No. 48 given in our on-line results, titled "High Performance Computing", one of the references is classified into the class of *microcomputers-architecture* which is considered a misclassification (i.e. an FP). However 2 out of 4 LCSHs assigned to this reference, which are *parallel processing (electronic computers)* and *supercomputers*, are quite relevant to the core subject of this document.
- In a majority of cases, classes assigned to a document can be grouped according to the higher level classes that they descend from. Also LCSHs are usually composed of a combination of different terms (e.g. *Computer networks -- Security measures*) and in many cases some of the LCSHs assigned to a document share the same first term. For example, in case of document No. 8, the subject headings *computer networks - security measures* and *computer networks* each appear once. Since these LCSHs share the same starting term, we can replace both of them with the heading *computer networks* and assign it the accumulated weights of the two original headings. This feature of the LCSHs and the hierarchical structure of the DDC allow us to apply an appropriate tradeoff between the recall and precision of our system according to users' requirements.

6 Conclusions and Future Work

In this paper, we looked at the problem of Automatic Text Classification from the perspective of researchers in the library science community. We specifically highlighted the potential application of controlled vocabularies, which have been initially developed for indexing and organizing physical library holdings, in the development

of ATC systems for organizing E-documents in digital libraries and archives. To do so, we first highlighted some of the up-to-date research work in this field and categorized associated ATC systems into two categories; ML-based systems and string matching-based systems, according to the approaches they have taken in leveraging library classification resources. We then proposed a third category of ATC systems based on a new route for leveraging library classification systems and resources, which we referred to as the Bibliography Based ATC (BB-ATC) approach. The proposed approach solely relies on the available classification metadata of publications referenced/cited in an E-document to classify it according to the DDC scheme. Unlike ML-based ATC systems, the new method does not require any training data or bootstrapping mechanism to be deployed. In order to demonstrate and evaluate the performance of proposed method, we developed a prototype ATC system for automatic classification of syllabus documents taken from the Irish National Syllabus Repository. The developed ATC system was evaluated with one hundred syllabus documents and the classification results were presented and analyzed with the aim of quantifying the prediction performance of the system and influencing factors. We reported micro-average values of 0.917, 0.889, and 0.902 for the precision, recall, and F1 performance measures of our system, respectively, and provided a relative comparison between the performance of our system and those of similar reported systems.

Based on above, we believe that we have developed a new robust method for ATC, which offers prediction performance that compares favorably to most other similar systems and outperforms many of them. As for future plans, we have identified a number of issues to be addressed, particularly with regards to enhancing the prediction performance of our method/system:

- Increasing the number of library catalogues that are queried by the catalogue-search unit. This would increase the number of references in the test documents with available classification metadata and, consequently, improve the system's performance in terms of recall. In order to do this, we are currently examining the use of a union catalogue, OCLC's WorldCat [23], which allows users to query the catalogues of 70,000 libraries around the world simultaneously.
- The Information Extractor component of our prototype classification system assumes that each reference entry includes an ISBN/ISSN number. Although this assumption holds for the documents that we tested the system with, in most cases citations in documents only include such information as title, author(s), and publisher's name. In order for our classifier to handle a reference with no ISBN/ISSN number, we need to extend the ability of its Information Extractor component such that it can locate and extract different segments of each reference entry and use them in the search queries submitted to the library catalogues instead of the ISBN/ISSN numbers. To achieve this, we are integrating an open source package called ParsCit [24] into the new version of BB-ATC system which can extract and segment a wide range of bibliographic entries.
- Koch et al. [25] studied users' navigation behaviors in a large web service, Renardus, by means of log analysis. Their study shows Directory-style of browsing in the DDC-based browsing structure to be clearly the dominant activity (60%) in Renardus. Conducting a similar study on users' navigation behaviors in Irish National Syllabus Repository could further justify the application of proposed method in improving information retrieval effectiveness of digital libraries.

- As mentioned in section two, libraries catalogue a wide range of publications including conference proceedings and journals. Therefore, one of the applications of the BB-ATC method is in organizing scientific literature digital libraries and repositories. We are currently working on an enhanced version of BB-ATC to classify about 700,000 scientific papers archived by CiteSeer [26] project.

References

[1] Avancini, H., Rauber, A., Sebastiani, F.: Organizing Digital Libraries by Automated Text Categorization. In: International Conference on Digital Libraries, ICDL 2004, New Delhi, India (2004)

[2] Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34(1), 1–47 (2002)

[3] Golub, K.: Automated subject classification of textual Web pages, based on a controlled vocabulary: Challenges and recommendations. New Review of Hypermedia and Multimedia 12(1), 11–27 (2006)

[4] Yi, K.: Automated Text Classification Using Library Classification Schemes: Trends, Issues, and Challenges. In: International Cataloguing and Bibliographic Control (ICBC), vol. 36(4) (2007)

[5] Dewey, M.: Dewey Decimal Classification (DDC) OCLC Online Computer Library Center (1876), http://www.oclc.org/us/en/dewey (cited January 2008)

[6] Putnam, H.: Library of Congress Classification (LCC) Library of Congress, Cataloging Policy and Support Office (1897), http://www.loc.gov/catdir/cpso/lcc.html (cited January 2008)

[7] Scorpion, OCLC Online Computer Library Center, Inc. (2002), http://www.oclc.org/research/software/scorpion/default.htm

[8] Larson, R.R.: Experiments in automatic Library of Congress Classification. Journal of the American Society for Information Science 43(2), 130–148 (1992)

[9] Jenkins, C., Jackson, M., Burden, P., Wallis, J.: Automatic classification of Web resources using Java and Dewey Decimal Classification. Computer Networks and ISDN Systems 30(1-7), 646–648 (1998)

[10] Dolin, R., Agrawal, D., Abbadi, E.E.: Scalable collection summarization and selection. In: Proceedings of the fourth ACM conference on Digital libraries, Berkeley, California, United States (1999)

[11] Chung, Y.M., Noh, Y.-H.: Developing a specialized directory system by automatically classifying Web documents. Journal of Information Science 29(2), 117–126 (2003)

[12] Pong, J.Y.-H., Kwok, R.C.-W., Lau, R.Y.-K., Hao, J.-X., Wong, P.C.-C.: A comparative study of two automatic document classification methods in a library setting. Journal of Information Science 34(2), 213–230 (2008)

[13] Frank, E., Paynter, G.W.: 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 (2004)

[14] Joorabchi, A., Mahdi, A.E.: A New Method for Bootstrapping an Automatic Text Classification System Utilizing Public Library Resources. In: Proceedings of the 19th Irish Conference on Artificial Intelligence and Cognitive Science, Cork, Ireland (August 2008)

[15] Sen, P., Namata, G.M., Bilgic, M., Getoor, L., Gallagher, B., Eliassi-Rad, T.: Collective Classification in Network Data. Technical Report CS-TR-4905, University of Maryland, College Park (2008), http://hdl.handle.net/1903/7546

[16] Joorabchi, A., Mahdi, A.E.: Development of a national syllabus repository for higher education in ireland. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 197–208. Springer, Heidelberg (2008)

[17] OpenOffice.org 2.0, sponsored by Sun Microsystems Inc., released under the open source LGPL licence (2007), http://www.openoffice.org/

[18] Xpdf 3.02, Glyph & Cog, LLC., Released under the open source GPL licence (2007), http://www.foolabs.com/xpdf/

[19] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, US (July 2002)

[20] Z39.50, International Standard Maintenance Agency - Library of Congress Network Development and MARC Standards Office, 2.0 (1992), http://www.loc.gov/z3950/agency/

[21] MARC standards. Library of Congress Network Development and MARC Standards Office (1999), http://www.loc.gov/marc/

[22] ISCED. International Standard Classification of Education -1997 version (ISCED 1997) (UNESCO (1997), http://www.uis.unesco.org (cited July 2008)

[23] WorldCat (Online Computer Library Center (OCLC) (2001)(2008), http://www.oclc.org/worldcat/default.htm (cited January 2008)

[24] Councill, I.G., Giles, C.L., Kan, M.-Y.: ParsCit: An open-source CRF reference string parsing package. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2008), Marrakesh, Morrocco (May 2008)

[25] Traugott, K., Anders, A., Koraljka, G.: Browsing and searching behavior in the renardus web service a study based on log analysis. In: Proceedings of the Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries, Tuscon, AZ, USA. ACM Press, New York (2004)

[26] Giles, C.L., Kurt, D.B., Steve, L.: CiteSeer: an automatic citation indexing system. In: Proceedings of the third ACM conference on Digital libraries, Pittsburgh, USA (1998)