Article

JIS

Journal of Information Science
XX (X) pp. 1-17

DOI: 10.1177/016555150000000
jis.sagepub.com
SAGE

An Unsupervised Approach to Automatic Classification of Scientific Literature Utilising Bibliographic Metadata

Arash Joorabchi and Abdulhussain E. Mahdi
Department of Electronic and Computer Engineering, University of Limerick, Limerick, Republic of Ireland

Abstract
This article describes an unsupervised approach for automatic classification of scientific literature archived in digital libraries and repositories according to a standard library classification scheme. The method is based on identifying all the references cited in the document to be classified and, using the subject classification metadata of extracted references as catalogued in existing conventional libraries, inferring the most probable class for the document itself with the help of a weighting mechanism. We have demonstrated the application of the proposed method and assessed its performance by developing a prototype software system for automatic classification of scientific documents according to the Dewey Decimal Classification (DDC) scheme. A dataset of one thousand research articles, papers, and reports from a well-known scientific digital library, CiteSeer, were used to evaluate the classification performance of the system. Detailed results of this experiment are presented and discussed.

Keywords
Digital library organisation; scientific literature classification; library classification schemes; citation networks; Dewey Decimal Classification (DDC); library Online Public Access Catalogues (OPACs); WorldCat; GBS

1. Introduction

Scientific digital libraries and repositories are a fast-growing concept within research and academic communities. The main aim of these services is to facilitate effective dissemination of research output among researchers by providing efficient centralised access points to large collections of research data in electronic format, mainly available in form of articles, papers, technical reports, thesis, and dissertations. Metadata, generally defined as data about data, plays a critical rule in digital libraries by providing structured data about characteristics of unstructured data resources. It can significantly improve the accessibility of resources by helping to describe, locate, and retrieve them efficiently. Hence, utilising data mining and knowledge discovery techniques to create, enrich, and harvest metadata has been one of the main efforts of researchers working in the field of digital libraries. The focus of this work is on a specific type of metadata called classification metadata (a.k.a. subject metadata) in scientific digital libraries, aimed at identifying the content subject of archived resources according to a standard classification scheme or taxonomy.

Medium to large-scale digital libraries contain tens to hundreds of thousands of items, and therefore require advanced querying and information retrieval techniques to facilitate precision search and discovery of archival materials. In order to deliver highly relevant search results, we need to go beyond the traditional keyword-based search techniques which usually yield a large volume of indiscriminant search results irrespective of their content. Subject classification of materials in digital libraries according to a standard scheme could improve the accuracy of information retrieval significantly and allows users to browse the collection by subject [1]. However, manual subject classification of documents is a tedious and time-consuming task which requires an expert cataloguer in each knowledge domain

Corresponding author:
Abdulhussain E. Mahdi, Department of Electronic and Computer Engineering, University of Limerick, Limerick, Republic of Ireland. Email: Hussain.Mahdi@ul.ie

represented in the collection, and therefore deemed impractical in many cases due to the sheer volume of new materials published on daily basis. For example, reportedly the number of new scientific publications in the field of biomedical science exceeds 1800 a day [2]. Motivated by the ever-increasing number of e-documents and the high cost of manual classification, Automatic Text Classification/Categorisation (ATC) - the automatic assignment of natural language text documents to one or more predefined classes/categories according to their contents - has become one of the key methods to enhance the information retrieval and knowledge management of digital textual collections.

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 labelled 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 application of ML algorithms in ATC, including the widely used Bayesian Model, *k*-Nearest Neighbour, and Support Vector Machine, is given in [3]. In general, an ML-based ATC 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, to name a few.

However, although a considerable success has been achieved in above listed applications, the prediction accuracy of ML-based ATC systems depends on a variety of factors, and no single ATC algorithm is adequate for all purposes. For example, it is commonly observed that as the number of classes in classification schemes increases, the prediction accuracy of ML algorithms decreases. This limitation of ML-based ATC systems becomes much more significant in case of scientific digital libraries where the classification schemes used could contain thousands of classes. Furthermore, the quality and quantity of the training dataset used to train the classification function has a decisive effect on the performance of ML-based ATC algorithms. However, in many cases, there is little or no training data available. consequently, over the past decade, most efforts of the ATC 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 labelled documents required to accurately train the classifiers.

On the other hand, as Golub [4], Yi [5], and Markey [6] 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 which have been developed and used for manual classification of holdings in conventional libraries. A library classification system is a coding system for organising library materials according to their subjects with the aim of simplifying subject browsing. Library classification systems are used by expert library cataloguers to classify books and other materials (e.g., serials, audiovisual materials, computer files, maps, manuscripts, realia) in conventional libraries. The two most widely used classification systems in libraries around the world today are the Dewey Decimal Classification (DDC) [7] and the Library of Congress Classification (LCC) [8], which since their introduction in the late 18th century have undergone numerous revisions and updates. A promising avenue for the application of this approach is the automatic classification of resources archived in digital libraries, where using standard library classification schemes is a natural and usually most suitable choice because of the similarities between conventional and digital libraries. Another application of this approach is in the classification of web pages, where due to their subject diversity, their proper and accurate labelling 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 virtually cover all categories and branches of human knowledge. In general, ATC systems that have been developed based on the above library science approach can be divided into two main categories:

1. String matching-based systems: these systems do not rely on 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 [9] by the Online Computer Library Centre (OCLC) [10]. 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 [11] in early 90's, who built normalised 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 another example of these systems see [12].

2. ML-based systems: these systems utilize ML algorithms to classify e-documents according to library classification schemes such as the DDC and the LCC. They represent a relatively unexplored trend which aims to combine the power of ML-based ATC algorithms with the enormous intellectual effort that has already been put into developing library classification systems over the last century. Chung and Noh [13] built a specialised web directory for the field of economics by classifying web pages into 757 sub-categories of economics category in the DDC scheme using *k*-NN algorithm. Pong et al. [14] 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 (NB) algorithms and compared the results. Frank and Paynter [15] used the linear SVM algorithm to classify over 20,000 scholarly Internet resources based on the LCC scheme. Wang [16] used both NB and SVM algorithms to classify a bibliographic dataset according to the DDC scheme and compared the results.

Golub et al. [17] have done an objective performance comparison between the string matching-based approach and the ML-based approach. The results of this study shows that the ML-based approach outperforms the string matching-based approach by a large margin. It also shows that combing the two approaches does not result in improved performance. These findings make the ML-based approach superior to the string matching-based approach. However, as discussed in [16], the large-scale and complexities of library classification schemes impose great obstacles on popular supervised ML-based classification algorithms (such as NB and SVM) and prevent them from reaching the high classification performances that these classifiers have reportedly achieved on standard benchmark datasets. These obstacles include: (a) deep hierarchy, where the hierarchical tree can go as deep as more than twenty levels; (b) skewed data distribution, where the great majority of training instances belong to a small number of classes; and (c) data sparseness, where there is a substantial number of classes which only have a few training instances, not sufficient for creating an accurate classification model.

In this work, we propose a new approach for automatic classification of scientific literature according to library classification schemes, with the aim of providing an easy-to-implement and efficient alternative to the ML-based approach for practitioners in the digital library community. Our approach, named Bibliography Based ATC (BB-ATC), does not require any training instances and is solely based on mining and utilizing the citation networks among scientific documents. The unsupervised nature of this approach allows practitioners to develop effective ATC systems for scientific digital libraries without encountering the obstacles mentioned above in relation to training data, which are associated with the ML-based approach. We demonstrate and evaluate the application of the BB-ATC approach in the automatic generation of subject classification metadata for documents archived in scientific digital libraries.

The rest of the paper is organised as follows: Section 2 introduces and provides an outline of the BB-ATC approach. Section 3 describes the implementation details of a prototype ATC system developed based on the BB-ATC approach in order to demonstrate its viability and evaluate its performance in organising a scientific digital library. Section 4 describes the evaluation process and presents its results. This is followed by Section 5 which analyses presented results and highlights some of the main factors affecting the performance of our method. Section 6 provides a conclusion along with a summary account of planned future work.

2. Introducing the BB-ATC Approach

A considerable amount of documents have some form of linkage to other documents. For example, it is a common practice in scientific documents to cite related papers, articles, and books. It is also common practice for documented law cases to refer to other cases, patents to refer to other patents, and webpages to have links to other webpages. Leveraging these networks of citations/links among documents has opened a new route for the development of ATC systems, known as collective classification [18]. Our proposed BB-ATC approach falls into this route, and aims to develop a new trend of effective ATC systems that are based on leveraging:

1. The intellectual work that has been put into developing and maintaining extensive resources and systems for classifying and organising the vast amount of materials archived in conventional libraries.
2. The intellectual effort of expert library cataloguers who have used the above classification resources and systems to manually classify and index millions of books and other materials in libraries around the world over the last century.

With the assumption that the majority of materials, such as books and journals, cited in a scientific document belong to the same or closely relevant classification category(ies) as that of the citing document, we can classify the citing

document based on the class(es) of its references as identified in existing conventional library catalogues. The proposed BB-ATC approach is based on automating this process using three main steps:

1. Identifying and extracting references in the document to be classified.
2. Searching for and retrieving the subject classification metadata of referenced materials from the online public access catalogues (OPACs) of conventional libraries.
3. Inferring and allocating a class(es) to the document based on the retrieved subject classification metadata of referenced materials with the help of a weighting mechanism.

This method of classification is applicable to any document that cites one or more published materials catalogued in at least one of the OPACs searched by the system. Examples of such documents include books, conference and journal articles, learning and teaching materials (e.g., syllabi and lecture notes), theses, and dissertations. In [19] the authors have described an ATC system designed and developed for automatic classification of electronic syllabus documents based on an early version of the BB-ATC method proposed here. Also, in [20] we have applied the underlying idea of the BB-ATC approach to the problem of automatic keyphrase indexing of scientific documents which could be viewed as a multi-label text classification problem. It should be noted here that the studies on the application of citation networks in automatic analysis of scientific documents go back to as early as 1955 [21]; and there is considerable recent literature on the use of citations networks to improve the search and retrieval of scholarly publications. For example, Aljaber et al. [22] showed that using citation contexts can provide relevant synonymous and related vocabulary which help increase the effectiveness of the bag-of-words representation used for clustering related scientific texts. Cao and Gao [23] showed that incorporating citation links data improves the accuracy of their ML-based system for classifying scientific documents. A series of work done by Bradshaw et al. [24, 25] and Ritchie et al. [26-28] demonstrated that using index terms from cited documents can optimize the full-text indexing and searching of scientific literature. However, to the best of our knowledge, the use of citation networks has only been studied in the context of supervised ML-based methods and/or ad hoc classification schemes (e.g., [23]).

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 DDC and the LCC schemes were considered as candidate classification schemes, due to their wide use and subject coverage. However, we eventually adopted the DDC for two main reasons:

1. The majority of libraries around the world use the DDC scheme, and therefore the number of items classified according to the DDC is much greater than those classified according to the LCC. This makes the DDC a better choice for our method which is based on mining and utilising the library assigned subject classification metadata of publications referenced in the document to be classified.
2. The DDC has a semantically hierarchical structure, whereas the LCC usually leans toward alphabetic or geographic sub-arrangements. The hierarchical structure of the DDC allows the development of effective GUI interfaces that enable users to easily browse and navigate the scheme to find the categories of their interest without requiring an extensive prior knowledge of the classification scheme or its notational representation [29].

The DDC scheme, currently in its 22nd version, is over a hundred years old and undergoes periodical updates and revisions. According to a recent study [30], DDC 22 contains ≈48,000 unique classes. The first, second, and third levels of the hierarchy contain 10, 100, and 1000 classes, respectively. It is worth mentioning here that the Universal Decimal Classification (UDC) [31] scheme, which is adapted from the DDC, has a number of advantages over the DDC (e.g., more faceted). There have been a number of efforts to use the UDC in automatic classification of electronic documents (e.g., see [32]), however, since the number of library collections indexed by the UDC is considerably lower than those indexed by the DDC, the latter was deemed more suitable for our method.

3. Implementation and Functionality

In order to demonstrate the application of the proposed BB-ATC method for automatic generation of subject classification metadata in scientific digital libraries, we have developed a prototype ATC system for categorising the scientific documents archived in the CiteSeer digital library [33] according to the DDC scheme. CiteSeer is a scientific literature digital library focusing primarily on the literature in computer science and information technology, and it contains over one million documents. We chose CiteSeer as our experimental platform for two main reasons:

1. CiteSeer is a well-known scientific digital library among the information science and digital library research communities and has been the subject of various studies in the areas of information retrieval and digital libraries.

2. It is an open-access and open-source research project providing full access to all of its resources including: metadata records, archived items, and software source codes.

Figure 1 shows an overview of our developed ATC system. As illustrated, the system is effectively a metadata generator comprising a pre-processing, a data mining, and an inferring unit. The complete collection of CiteSeer's metadata records is freely available on the project's website[1] in the form of dump files. CiteSeer metadata records come in two different types: Open Access Initiative (OAI) records in Dublin Core XML format and bibliographic records in BibTex format. These two types of metadata records associated to each archived document contain a wide range of metadata about the document such as: type, title, authors, abstract, references, publishing date, publisher, source URL, format, language, etc. In order to easily access this large collection of metadata records we first developed a small software component to normalise and convert the CiteSeer BibTex records into XML format. Then the CiteSeer OAI and BibTex records in XML format were loaded into a native XML database called eXist-DB [34] which supports XML query languages, Xquery and Xpath, and facilitates efficient search and retrieval of CiteSeer metadata records.

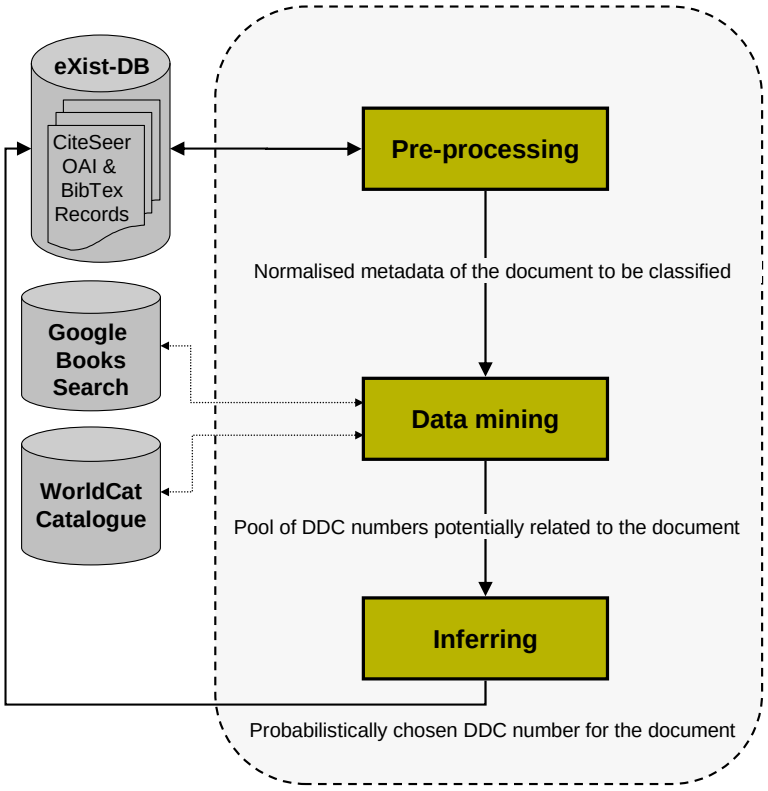

Figure 1. Overview of the prototype ATC system

The initial task of the pre-processing unit is to select a document from the CiteSeer archive for classification and retrieve its metadata from the CiteSeer metadata database for further processing. The selection can be sequential, random, or based on some criteria, such as publishing date, number of references, format, etc. Once a document is selected and its metadata is retrieved, the pre-processing unit compiles a list of titles of all the publications referenced in the document, such as articles, books, reports, etc., as per the list of references provided in the CiteSeer OAI metadata record of the document. The retrieved metadata of the document along with its list of references are then passed to the data mining unit.

The task of the data mining unit is two fold. In the first stage, it uses the Google Books Search (GBS) engine [35] to compile a list of publications that either cite the document to be classified or one of its references. This is done by submitting a number of URL queries to the GBS engine in the following format:

http://www.google.com/books/feeds/volumes?**max-results=*20***&q=%22**[*title*]**%2C%22

[1] http://citeseer.ist.psu.edu/

For the first query, the variable *title* in above format is set to the title of the document to be classified, and in the subsequent queries the titles of the references in the document are used consecutively. The parameter *max-results* limits the number of returned matching results to twenty items. This parameter is set empirically to balance the bias in the search results in terms of the number of returned matching publications for different queries. The returned result for each query is an XML file containing the metadata records of matching publications and each record contains a set of metadata elements such as: title, authors, ISBN, etc. At this point, we have a pool of metadata records for the publications that either cite the document to be classified or one of its references. In order to utilise the gathered metadata for inferring the DDC class of the document, we first need to discover the DDC classification numbers of the publications in the pool. This is achieved by the second stage of the data mining process, where the corresponding DDC numbers of publications in the pool are retrieved from the OCLC's WorldCat [36] database. WorldCat is a union catalogue of about 70,000 conventional libraries around the world. The data mining unit performs this task in two steps. First, it processes the metadata records of the publications in the pool to extract their corresponding ISBNs. These ISBNs are then used as unique identifiers for the publications to query the WorldCat database for their corresponding metadata records. The latter process is done by submitting the following URL query to the WorldCat Search API [37] per each ISBN:

http://www.worldcat.org/webservices/catalog/content/isbn/**[*ISBN*]**%3Fwskey%3D**[*key*]**%3Dfull

where, the parameter *key* is a unique identifier assigned to each developer to access the API. The returned result for each query is and XML file containing the full bibliographic record of the publication in MARC 21 XML format [38]. Along with other metadata elements, this record contains a DDC classification number assigned to the publication by a professional library cataloguer in one of the 70,000 libraries that have merged their catalogue into the WorldCat catalogue. Figure 2 illustrates the two-folded data mining process described above.

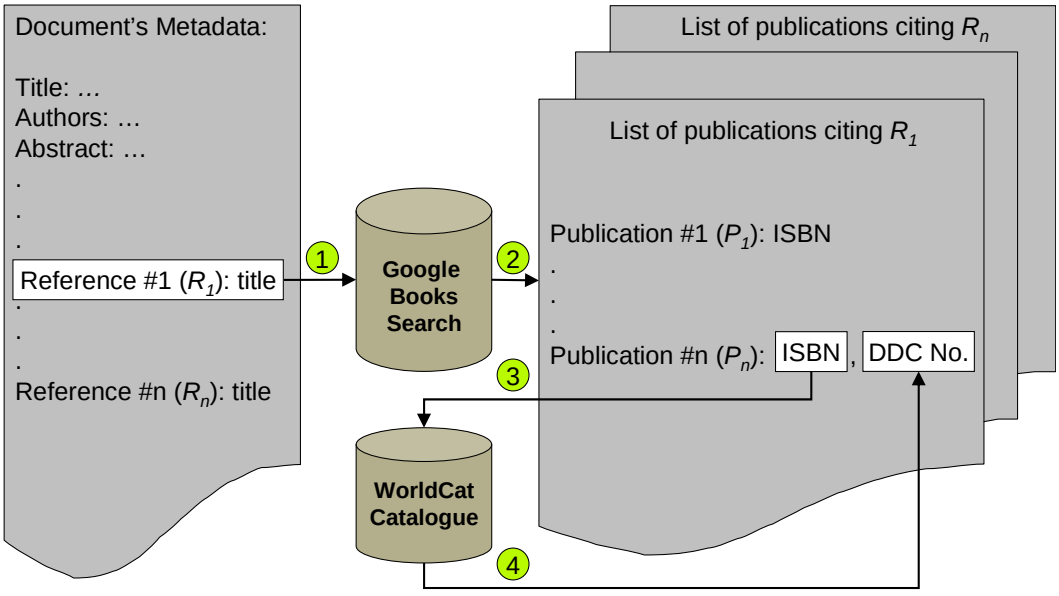

Figure 2. Data mining process

The task of the inferring unit is to analyse the pool of metadata gathered by the data mining unit, which contains the DDC numbers potentially related to the document to be classified, and select a DDC number from the pool which is most probable to represent the document's core subject. The inference process is based on a weighting method designed to assign a relevance probability score to each unique DDC number in the pool according to its frequency distribution.

Initially, the weighting method assigns each unique DDC number in the pool three different weights: un-normalised local frequency, normalised local frequency, and global frequency. Each of these weights is designed to measure the relevance probability of a given DDC number in the pool in relation to the document from a unique perspective. We describe these weights and details of the inferring process in the course of the following example which gives a sequential account of how the proposed BB-ATC method is used to classify a sample document from the CiteSeer archive. The document used in this example is a research paper entitled "Statistical Learning, Localization, and Identification of Objects". The core subject of the document is AI-based computer vision and, therefore, it should be

classified into the DDC class "Computer science, information & general works\Computer science, knowledge & systems\Special computer methods\Artificial intelligence\Computer vision" represented by the DDC number 006.37. The classification process of this sample document would be as follows:

3.1. Pre-processing

The pre-processing unit retrieves the corresponding metadata records for the document to be classified from the CiteSeer metadata database. Table 1 shows some of the retrieved metadata for the sample document.

Table 1. Sample document's metadata

Metadata field	Value
identifier	oai:CiteSeerPSU:52
datestamp	1996-08-06
dc:title	Statistical Learning, Localization, and Identification of Objects
dc:description	This work describes a statistical approach to deal with learning and recognition problems in the field of computer vision…
dc:publisher	Unknown
dc:contributor	The Pennsylvania State University CiteSeer Archives
dc:format	Ps
dc:identifier	http://citeseer.ist.psu.edu/52.html
dc:source	http://www5.informatik.uni-erlangen.de/TeX/Literatur/ps-dir/1995/Hornegger95:SLL.ps.gz
dc:language	En
oai_citeseer:relation type="References"	<oai_citeseer:uri>oai:CiteSeerPSU:112462</oai_citeseer:uri>

3.2. Data mining

As described earlier, this process involves compiling a list of publications that either cite the document to be classified or one of its references, and discovering their corresponding DDC numbers. As the last row of Table 1 shows, the document under classification either has only one reference, or the CiteSeer's citation extraction unit, ParsCit [39], which is responsible for extracting citations from the archived documents, has only managed to extract one of the references successfully. Therefore, the title of the document to be classified and the title of its single successfully extracted reference are the only available clues that can be used for mining a list of DDC numbers potentially relevant to the document. Table 2 shows the metadata gathered by the data mining unit for the publications that cite one of these two titles.

Table 2. Data mining results for the sample document

Publications citing the document to be classified titled: "Statistical Learning, Localization, and Identification of Objects"							
ISBN	**DDC No.**	**ISBN**	**DDC No.**	**ISBN**	**DDC No.**	**ISBN**	**DDC No.**
0123797721	006.37	3540650806	006.3	0818681845	621.367	0769501648	Null
0123797772	006.37	3540629092	006.42	3540639314	621.367	1558605835	Null
3540646132	006.37	3540634606	006.42	0792378504	621.367	0780399781	Null
0780350987	006.37	389838019X	005.118	3540250468	629.8932		
Publications citing the document's reference titled: "Learning Object Recognition Models from Images"							
ISBN	**DDC No.**	**ISBN**	**DDC No.**	**ISBN**	**DDC No.**	**ISBN**	**DDC No.**
3540617507	006.37	1586032577	006.3	389838019X	005.118	0120147734	537.56
0195095227	006.37	3540282262	006.3	1848002785	621.367	0780399773	Null
3540667229	006.37	3540634606	006.42	3540433996	629.892		
3540404988	006.37	3540636366	006.7	0818638702	621.399		

3.3. Inferring

The inference process starts by deriving the un-normalised local frequency, normalised local frequency, and global frequency weights for each unique DDC number in the pool, as per the following:

- The un-normalised Local Frequency (ULF) of a given DDC number, DDC_i, is defined as the total summation of its frequencies in each of the search result sets, R_j, where $j = \{1,\ldots, m\}$ with m being the total number of search result sets:

$$ULF(DDC_i) = \sum_{j=1}^{m} Freq(DDC_{i,j}) \quad (1)$$

  The function $Freq(DDC_{i,j})$ returns the number of times that the DDC number, DDC_i, appears in the search result set R_j. For a given DDC number, DDC_i, which appears in the pool of search results at least once, $ULF(DDC_i)$ is an integer number greater than or equal to 1. For example, the result of data mining process for the sample document appearing in Table 2 shows that there are 12 publications with valid DDC numbers (not null) that cite the document to be classified; and there are 13 publications with valid DDC numbers that cite the document's only reference. Among this total of 25 publications, 8 are assigned the DDC number "006.37", and therefore the ULF value for this DDC number is equal to 8.

- In order to prevent a DDC number from unjustifiably biasing the inference result due to its overwhelming high frequency in a single or small number of search result sets, we have adopted a second weight called Normalised Local Frequency (NLF) defined as:

$$NLF(DDC_i) = \sum_{j=1}^{m} \frac{Freq(DDC_{i,j})}{|R_j|} \quad (2)$$

  where, $|R_j|$ represents the total number of valid DDC numbers in the search result set R_j. For a given DDC number, DDC_i, which appears in the pool of search results at least once, $NLF(DDC_i)$ is a positive real number greater than 0. For example, using the sample data given in Table 2, the NLF value for the DDC number "006.37" is (4/12) + (4/13) = 0.64.

- The third weight, Global Fequency (GF), aims to reflect how common a given DDC number is among all the search result sets irrespective of its frequency inside individual search result sets. The GF for a given DDC number, DDC_i, is defined as the total number of search result sets in which DDC_i appears once or more:

$$GF(DDC_i) = \sum_{j=1}^{m} \left[DDC_i \in R_j\right] \quad (3)$$

  where, $[DDC_i \in R_j]$ returns 1 if DDC_i appears in the search result set R_j at least once, and returns 0 otherwise. For a given DDC number, DDC_i, which appears in the pool of search results at least once, $GF(DDC_i)$ is a positive integer number, with a minimum value of 1 and a maximum value of m, with m being the total number of search result sets. Again, using the sample data given in Table 2, the DDC number "006.37", for example, appears in both R_1 and R_2 search result sets, and therefore its GF is equal to 2.

Having computed the ULF, NLF, and GF weights for a given DDC number, DDC_i, the formula in Equation 4 is used to derive a single Combined Weight (CW) for it:

$$CW(DDC_i) = GF(DDC_i) \times NLF(DDC_i) \times ULF(DDC_i)^{\frac{depth(DDC_i)}{10}+1} \quad (4)$$

where, $depth(DDC_i)$ returns the vertical position of DDC_i in the classification hierarchy. The formulas for the ULF, NLF, and GF weights of a given DDC number in the pool and the CW formula used to derive a single combined weight from them, have been empirically deduced to give the best inference results based on an extensive analysis of a preliminary dataset. The results of this analysis indicated that the impact of ULF on CW should be kept to a minimum for the DDC numbers at the first level of the DDC hierarchy and it should gradually increase as the depth/level of the given DDC number increases in the hierarchy. The last part of Equation 4 incorporates this condition. Sticking to the DDC number "006.37" in our example and using the data of Table 2, the CW for this DDC number is computed as: $2 \times 0.64 \times 8^{(5/10)+1} = 29.01$.

After computing the above weights for all the DDC numbers in the pool, the inferring unit builds a classification hierarchy tree from all the DDC numbers in the pool and their corresponding weights. This tree is then automatically

inspected to find the most probabilistically relevant DDC number to the core subject of the document. The inferring unit uses Java Universal Network/Graph Framework (JUNG) [40], which is an open source software library for graph modelling, analysis, and visualisation, to build, crawl, and visualise the classification tree. For example, Figure 3 shows part of the classification tree built and visualised by the inferring unit for the sample document discussed in this section. In the figure, the automatically selected classification path leading to the final chosen DDC number (i.e., 006.37) for the document appears in red.

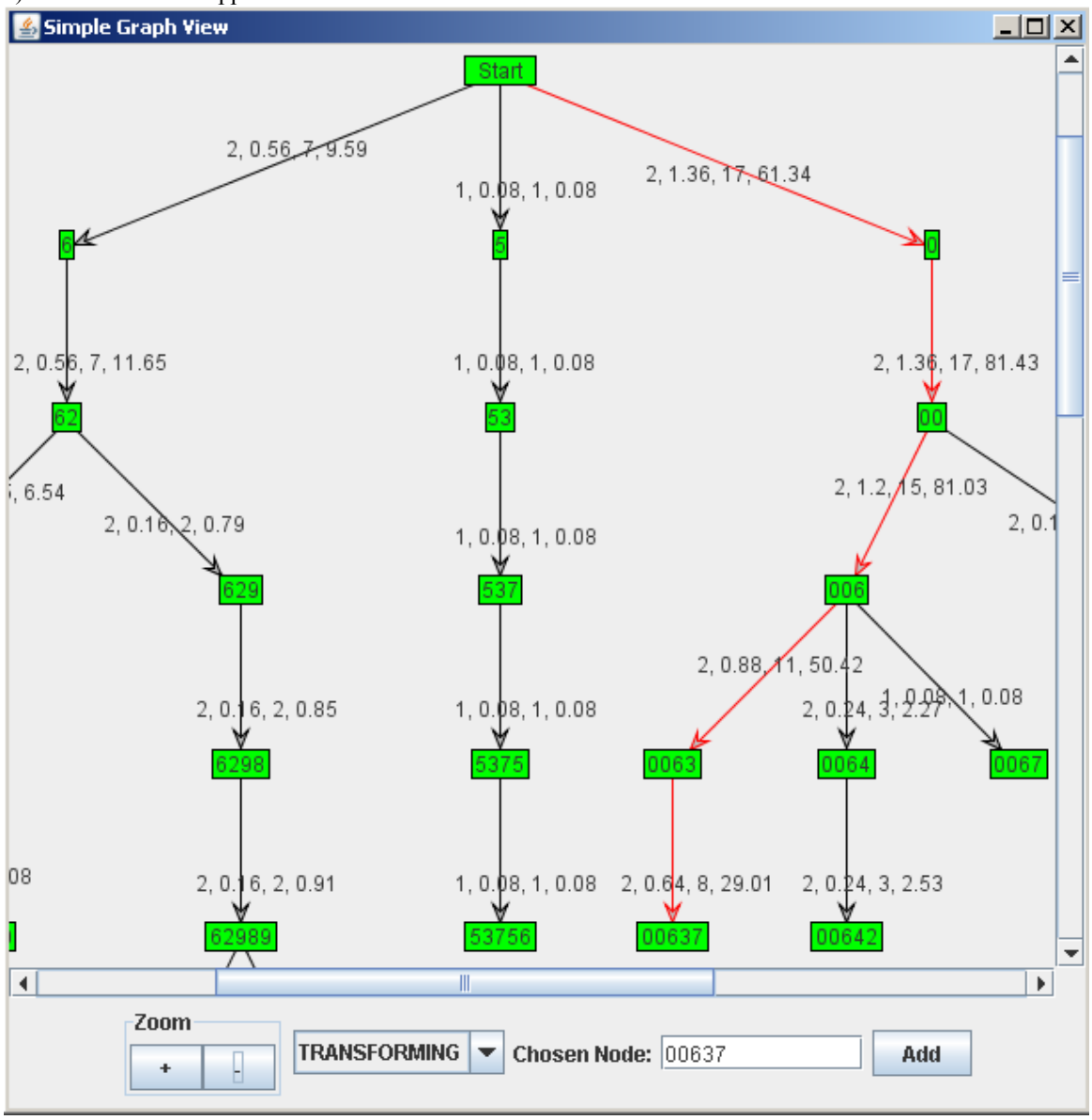

Figure 3.Sample visualised output of the inferring unit

As shown, each vertex/node in the tree represents a unique DDC number and is connected to its parent node by an edge/link labelled with its corresponding weights. In specific, the comma-separated values in the label of each link are the GF, NLF, ULF, and CW, respectively. The automatic crawling process aims to find the strongest path in the classification tree based on the CW values of the nodes. It starts from the root/start node and moves to the child node which has the largest CW value as the probabilistically selected most relevant DDC number to the document in the first level of the DDC classification hierarchy. The same selection criterion is then applied to the children of the selected node and so on until a node with no children (i.e. a leaf node) is reached. In cases where all the children of a chosen node have equal CW values, the CWs of its grandchildren are compared and the grandchild which has the largest CW value along with its parent node become selected. If all the grandchildren of the chosen node have equal CW values, then the decision will be based on the CWs of its great grandchildren and so on. In rare cases where this selection

criterion does not lead to a resolution and all of the descendents of the latest chosen node have equal CW values in their corresponding level of the classification hierarchy, the crawling process stops and the latest selected node becomes the final selected DDC number for the document.

During our preliminary experiments, we noticed some cases where there is a significant decrease in the CW value of a potentially chosen node in relation to its parent's CW value. In majority of studied cases, this sudden drop indicated that either the latest chosen node (i.e., the parent node of current potentially chosen node) is the most appropriate DDC number for the document or there is not enough evidence to confidently conclude otherwise. In these cases, the best policy is to stop the crawling process and output the latest confidently chosen node, i.e., the parent node, as the final selected DDC number for the document. This policy is incorporated into the inference process in the form of a thresholding mechanism which stops the crawling process if a potentially chosen node does not pass the criterion in Equation 5, and outputs the parent of that node as the final selected DDC number for the document.

$$CW(CN) \times depth(CN)^{children(PN)} > CW(PN) \quad (5)$$

where, *CW(CN)* is the *CW* value of the current potentially chosen node, *depth(CN)* is the depth of the current potentially chosen node in the DDC hierarchy, *children(PN)* is the number of the parent node's children (i.e., the number of the current potentially chosen node's siblings added by one), and *CW(PN)* is the *CW* value of the parent node.

The prototype BB-ATC system operates in two modes: unsupervised and semi-supervised/evaluation. In unsupervised mode, classification process of a document ends by adding its final chosen DDC number to its metadata record stored in the CiteSeer metadata database. In the semi-supervised mode, however, first, the built classification hierarchy tree and the inference result for the document are visualised and presented (as shown in Figure 3); and then the user is required to either confirm the DDC number suggested by the system for the document as the most appropriate, or enter the correct DDC number manually. Once the results are confirmed/corrected, both the DDC numbers chosen by the inferring unit and the user are added to the metadata record of the document stored in the CiteSeer metadata database. In parallel to that, when operating in evaluation mode, the system creates a HTML log file for each classified document containing its original metadata, data mining results, manual and automatic generated subject classification metadata, and an image of its visualised classification hierarchy tree for future analysis.

4. Evaluation & Experimental Results

Evaluating the performance of the developed prototype BB-ATC system was the most challenging and time consuming part of this work. To start with, the CiteSeer digital library, used as the test platform in this work, does not provide any subject classification metadata for its archived items. In fact, to the best of our knowledge, there exit no digital library of scientific literature which classifies its collection according to a standard library classification scheme, such as the DDC or the LCC. This fact, as discussed in Section 1, can be attributed to two main obstacles: the first is the high cost of manual classification, and the second is the inefficiency of common ML-based ATC systems to cope with the large scale and complexities of library classification schemes, containing thousands of classes. Therefore, in the absence of any suitable third-party test corpus, we had no option but to create our own.

To perform the evaluation, the pre-processing unit of the system was set to randomly retrieve the metadata records of one thousand documents from the CiteSeer metadata database to be automatically classified and manually examined by a group of five postgraduate students in our research group. The students were given access to the WebDewey[2] which is part of the Online Computer Library Centre (OCLC) [10] suite of cataloguing and metadata services and enables full browsing of the latest version of the DDC online. The students were first familiarised with the DDC scheme and its hierarchical nature, and then each were assigned a set of documents (as defined below) to examine and classify.

In order to measure the effect of the number of references successfully extracted from documents on the classification performance of our system, the pre-processing unit was set to randomly build the test corpus from five different subsets of documents grouped according to their number of references. Each subset is made up of 200 documents with equal number of references. The first subset contains the documents that have no references successfully extracted from them. The second, third, forth, and fifth subsets contain documents with 4, 8, 16, and 32 references, respectively. Also, we set the inferring unit of the system to work in the semi-supervised/evaluation mode, which requires the user to either verify or rectify the DDC number automatically assigned to the document, and logs all

[2] http://www.oclc.org/dewey/versions/webdewey

the data produced during the classification process of the document in a dedicated HTML log file, as explained previously in Section 3. The HTML log files for all of the 1000 test documents used in this experiment may be viewed online on our webpage[3].

We used the standard measures of Precision (*Pr*), Recall (*Re*), and *F1* to evaluate the classification performance of our system. Precision is the probability that a document predicted to be in category, c_i, truly belongs to this category. Recall is the probability that a document belonging to c_i is classified into this category. When a single performance measure is desired, the harmonic mean of precision and recall, F1, is quoted. Accordingly with respect to a given class c_i:

$$Pr(c_i) = \frac{\text{Number of correctly assigned class labels}}{\text{Total assigned}} = \frac{TP_i}{TP_i + FP_i} \quad (6)$$

$$Re(c_i) = \frac{\text{Number of correctly assigned class labels}}{\text{Total possible correct}} = \frac{TP_i}{TP_i + FN_i} \quad (7)$$

$$F1(c_i) = \frac{2Pr(c_i)Re(c_i)}{Pre(c_i) + Re(c_i)} \quad (8)$$

where, the *Pr*, *Re*, and *F1* are computed in terms of the labels *TP* (True Positive), *FP* (False Positive), and *FN* (False Negative) to evaluate the validity of a given class label *i* assigned to a given document *j*, such that:

- TP_i: refers to the cases when both the classifier and human cataloguer agree on assigning class label *i* to document *j*;
- FP_i: refers to the cases when the classifier has mistakenly (as judged by a human cataloguer) has assigned class label *i* to document *j*;
- FN_i: refers to the cases when the classifier has failed (as judged by a human cataloguer) to assign a correct class label *i* to document *j*.

Micro-average and macro-average are the two wildly used measures to evaluate the overall prediction performance of ATC systems. In micro-averaging, the above target performance measures (i.e. *Pr*, *Re*, and *F1*) are computed globally over all classes. Whereas, in macro-averaging, the performance measures are computed for each individual class locally and then the average over all classes is taken. Micro-averaging gives equal weight to each document, whereas, macro-averaging gives equal weight to each class. Due to the high subject sparsity of our test corpus, there is a substantial number of classes which contain only one or two documents, and that could result in biased performance measures if macro-averaging is used. Therefore, in order to obtain a true objective evaluation of the classification performance of our system, we adopted the micro-average measure which gives equal weight to each document regardless of its class. The overall micro-averaged precision, recall, and F1 for all the one thousand documents in the test corpus regardless of their number of references are 0.84, 0.78, and 0.81 respectively. In order to show the effect of the number of references in the documents on the classification performance of the system, Table 3 shows the micro-averaged performance measures for each of the five document subsets in the test corpus individually. It also shows the number of documents which truly belong to classes in each level of the DDC hierarchy and the respective performance measures achieved in each individual level.

As a common practice in developing a new ATC method or system, it is always desired to compare its performance with that of others. However, it was not possible for us to conduct a true objective comparison between the performance of our system and that of other reported ATC systems due to the following:

1. To the best of our knowledge there has been no previous attempt to automatically classify a collection of digital scientific literature according to a standard library classification scheme.
2. Unlike our system which utilizes the full DDC scheme, other relatively similar reported ATC systems, due to their limitations, either adopt only one of the main classes in the DDC/LCC along with its subclasses as their classification scheme, or use an abridged version of the DDC/LCC by limiting the depth of the classification hierarchy to second or third level.

[3] http://www.csn.ul.ie/~arash/BB-ATC1/HTML/index.html

3. Some of the similar works had reported the performance of their system using measures other than the standard performance measures of precision, recall, and F1 used in this work.

Table 3. Micro-averaged performance measures for each of the five document subsets in the test corpus

Subset	#References	DDC level	#Docs per level	Precision	Recall	F1
1	0	1	200	0.88	0.64	0.74
		2	200	0.85	0.62	0.72
		3	200	0.72	0.52	0.61
		4	200	0.68	0.48	0.56
		5	189	0.63	0.39	0.48
		6	69	0.44	0.41	0.42
		7	7	0.27	0.43	0.33
		8	2	0.20	0.50	0.29
		Overall	**200**	**0.72**	**0.52**	**0.61**
2	4	1	200	0.94	0.94	0.94
		2	200	0.93	0.93	0.93
		3	200	0.84	0.84	0.84
		4	200	0.83	0.82	0.82
		5	193	0.74	0.67	0.70
		6	69	0.68	0.58	0.63
		7	6	0.30	0.50	0.38
		8	3	0.60	1.00	0.75
		Overall	**200**	**0.84**	**0.82**	**0.83**
3	8	1	200	0.96	0.96	0.96
		2	200	0.94	0.94	0.94
		3	200	0.89	0.89	0.89
		4	200	0.82	0.81	0.81
		5	192	0.74	0.70	0.72
		6	99	0.73	0.63	0.67
		7	13	0.47	0.54	0.50
		8	6	0.29	0.33	0.31
		Overall	**200**	**0.84**	**0.83**	**0.84**
4	16	1	200	0.97	0.97	0.97
		2	200	0.95	0.95	0.95
		3	200	0.89	0.89	0.89
		4	200	0.86	0.86	0.86
		5	189	0.78	0.74	0.76
		6	86	0.73	0.66	0.70
		7	9	1.00	0.67	0.80
		8	4	1.00	0.50	0.67
		Overall	**200**	**0.88**	**0.86**	**0.87**
5	32	1	200	0.96	0.96	0.96
		2	200	0.94	0.94	0.94
		3	200	0.88	0.88	0.88
		4	200	0.87	0.86	0.86
		5	187	0.84	0.82	0.83
		6	71	0.80	0.85	0.82
		7	15	0.92	0.73	0.82
		8	5	1.00	0.40	0.57
		Overall	**200**	**0.89**	**0.88**	**0.89**

Despite above, it is possible to provide a relative comparison between the performance of our system and those of similar systems reported in the literature. For example, Pong and co-workers [14] used both NB and *k*-NN algorithms to classify 254 documents according to a refined version of the 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 [13] reported the development of a specialised 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 [19] we used an early version of the BB-ATC method to automatically classify a collection of 200 computer science related syllabus documents archived in the Irish national syllabus repository according to the full DDC scheme. The achieved micro-averaged performance measures of precision, recall, and F1 were 0.917, 0.889, and 0.902, respectively.

5. Discussion of Results

During the evaluation process, we manually examined the classification result for each individual document in the test corpus in order to validate the prediction performance of our system, as well as identifying factors that affect this performance. In doing this, we first read the abstract and/or the introduction section of the document and then examined its extracted references, data mining results, and visualized classification tree. In this section we summarise the findings of this examination process.

As expected, the number of references in the documents plays a critical role in the performance of our ATC system. This role can be observed in the overall performance results for the five document subsets in the test corpus, presented in Table 3. There is a considerable improvement in micro-averaged F1 score mounting to 23% when the number of references in the documents is increased from 0 to 4. In contrast, there is relatively small improvement of 0.5%, 3.5% and 1.6% in F1 score when the number of references is doubled to 8, 16 and 32, respectively. Based on this observation, we can conclude that although increasing the number of references in the documents generally results in improving the prediction performance of our system, for it to achieve an acceptable level of classification accuracy, it does not require the documents to have a large number of references, and with a modest number of references (4 in this case) it can yield an acceptable micro-averaged F1 score of 0.83.

Half of the documents in our test corpus have only 0 to 8 references each, and this has made a considerable negative impact on the overall performance scores of our system. In practice, the number of references in scientific documents depends on a number of factors, with the following being the most important ones: (a) type of the document. For example journal articles usually have more references than conference papers. Similarly, Electronic Theses and Dissertations (ETDs) tend to have more references than journal articles; (b) field of the document. For example, a study on the 2008 Journal Citation Reports (JCR) from the ISIWeb of Knowledge (WoK)[4] has shown that the average number of references per publication for 35 fields of science (with more than 3000 articles in each) widely ranges from a minimum of 18.5 for the field of "Mathematics" to a maximum of 51.6 for the field of "Cell Biology" [41]. Based on this we envisage the average number of references per document in scientific digital libraries such as CiteSeer to be no less than 16 per publication. Our prototype classification system has achieved an overall micro-averaged F1 score of 0.87 for subset no. 4 in the test corpus which contains 200 documents with exactly 16 references per document. Therefore, we envisage the BB-ATC method to achieve a similar performance in real-world applications.

Table 4 shows the percentages of the documents from the whole test corpus (regardless of their number of references) per each level of the DDC hierarchy and the corresponding performance measures achieved in each level. As can be seen in the table, 95% of all documents in the test corpus reach at least level 5 of the DDC hierarchy and a mean micro-averaged F1 score of 0.70 is achieved at this level, with 0.48 and 0.83 being the minimum and the maximum micro-averaged F1 scores achieved at this level, depending on the number of references in the documents. The data in Table 4 demonstrates a reverse relationship between the level of classification and the mean performance measures achieved. However, the percentage of documents in the test corpus whose subjects are very specific and require classification at levels 7 and 8 is limited to %5 and 2%, respectively. This in effect indicates that the relatively low performance scores achieved in these deep levels of the DDC hierarchy would not have a significant negative impact on the system's overall performance. For the readers who are not familiar with the DDC scheme and the

[4] http://thomsonreuters.com/products_services/science/science_products/a-z/isi_web_of_knowledge/

degrees of subject specialty in its different hierarchical levels, Figure 4 shows the ascendants and descendants of two computer science related DDC classes "artificial intelligence" and "computer pattern recognition" from levels 1 to 5. For example, the sample document used in Section 3 to illustrate the classification process belongs to the class "computer vision" highlighted in this figure.

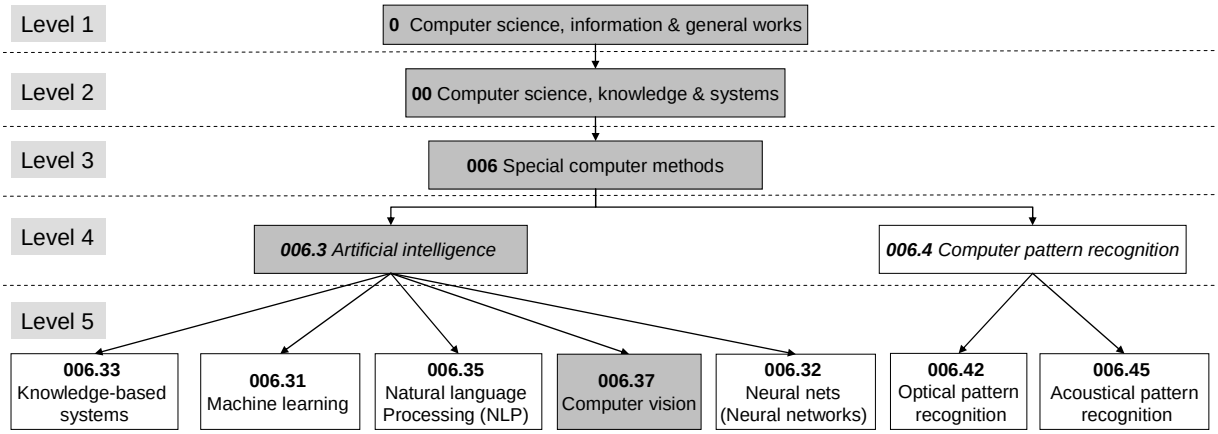

Figure 4. Ascendants and descendants of two sample classes in the DDC hierarchy from levels 1 to 5

Table 4. Distribution of test documents among the DDC hierarchical levels and corresponding performance measures achieved in each level

DDC Level	#Docs	Micro-Averaged Precision			Micro-Averaged Recall			Micro-Averaged F1		
		Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
1	1000	0.88	0.94	0.97	0.64	0.89	0.97	0.74	0.91	0.97
2	1000	0.85	0.92	0.95	0.62	0.87	0.95	0.72	0.89	0.95
3	1000	0.72	0.84	0.89	0.52	0.80	0.89	0.61	0.82	0.89
4	1000	0.68	0.81	0.87	0.48	0.77	0.86	0.56	0.79	0.86
5	950	0.63	0.75	0.84	0.39	0.66	0.82	0.48	0.70	0.83
6	394	0.44	0.68	0.80	0.41	0.63	0.85	0.42	0.65	0.82
7	50	0.27	0.59	1.00	0.43	0.57	0.73	0.33	0.58	0.82
8	20	0.20	0.62	1.00	0.33	0.55	1.00	0.29	0.58	0.75

Our prototype BB-ATC system was designed to carry out single-label classification, where each document is categorised into only a single class most relevant to the main subject of the document. During the evaluation process we encountered a considerable number of cases, where although the class assigned to a document did not represent the main/dominant subject of the document, it still revealed an important aspect of the document's content. However, since the designed aim of the system was to carry out single-label classification and therefore only one class could have been assigned to each document, these cases were considered as misclassifications during the evaluation process and, hence, reflected negatively on the performance scores of the system. As an example of these cases, consider a document discussing the application of wavelets in image processing categorised into the DDC class "Fourier and harmonic analysis" by the system. In this case, the class assigned to the document is quite relevant, however since the main subject of the document is image processing and only one class must be assigned to it, we consider this case a misclassification resulting in a False Positive (FP) for the class "Fourier and harmonic analysis" and a False Negative (FN) for the class "optical pattern recognition". This indicates that adopting a multi-label classification approach, where each document can be categorised into multiple classes, could potentially improve the systems classification performance scores and also result in richer classification metadata. However, this hypothesis remains to be tested in future experiments.

6. Conclusions and Future Work

In this article, we looked at the problem of automatic text classification from the perspective of researchers in the library science community. Specifically, we highlighted the potential application of controlled vocabularies, which were originally developed for indexing and organising conventional library holdings, in the development of ATC systems for subject classification of documents archived in scientific digital libraries and repositories. To do so, we first reviewed some of the up-to-date research works in this field and categorised them into two main groups of ML-based systems and string matching-based systems, according to their approaches to leveraging conventional library classification resources. We then proposed a third category of such ATC systems based on a new route for leveraging library classification systems and resources, which we refer to as the Bibliography Based ATC (BB-ATC) approach. Our proposed approach solely relies on the subject classification metadata of the publications citing either the document to be classified or one of its references, as catalogued in the OPACs of conventional libraries, in order to probabilistically infer the most appropriate class for the document. To demonstrate the application and evaluate the classification performance of the proposed BB-ATC approach, we developed a prototype ATC system for automatic classification of scientific literature archived in the CiteSeer digital library. The developed ATC system was evaluated using a test corpus of one thousand scientific documents and the classification results were presented and analysed with the aim of quantifying the prediction performance of the system and identifying the factors influencing its performance. We reported micro-averaged values of 0.84, 0.78, and 0.81 for the overall precision, recall, and F1 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 unsupervised approach to automatic classification of scientific literature in digital libraries according to standard library classification schemes, which yields a prediction performance competitive to that achieved by the ML-based approach and offers an effective alternative to digital library practitioners. As for future work, we have identified a number of enhancements that could potentially improve the prediction performance of our method:

- As discussed in section 3, in the first stage of the data mining process carried out by the data mining unit of our system, the GBS engine is used to gather the corresponding metadata of the publications that either cite the document to be classified or one of its references. GBS enables the full-text search of books, journals, and other materials that Google and its library and publisher partners scan, OCR, and index. In October 2009, Google announced that they had over 10 million items searchable through the GBS [42]. Google does not provide public access to the full content of the majority of these items due to copyright restrictions. However, the metadata record of each archived item includes a so called "word cloud" which contains a set of key terms that have been identified as statistically significant within the full textual content of the item. Figure 5 shows the Google Word Cloud (GWC) for a book titled: "Data mining: practical machine learning tools and techniques". The majority of these key terms are domain-specific, semantically rich, and directly related to the core subject of the book, and we have already proved their application in automatic keyphrase extraction from scientific documents [20]. These key terms could be used to measure the relevance of a publication, which cites either the document to be classified or one of its references, to the document. Thus, we are currently working on an enhanced version of the BB-ATC system which searches the content of the document to be classified for these key terms and based on their total number and frequency in the document derives a relevance weight, which measures the subject similarity of the citing publications to the document. Incorporating this new weight into the inference process should eliminate or at least limit the negative effect of a minor number of citing publications, whose main subject does not match the subject of the document to be classified. It should be noted here that we have empirically set the maximum number of GBS results retrieved per query to 20 in order to balance the bias in the search results in terms of the number of returned matching publications for different queries. Currently, we are objectively investigating the impact of the number of GBS results retrieved per query on the predication performance measures of the system.

Figure 5. A sample GWC from GBS database

- The number of references in the documents to be classified has a large impact on the prediction performance of the proposed method. The references are used as indicative clues which collectively point to the right class for the document and, therefore, the larger the number of the clues the more reliable and accurate the classification results. Based on this, we can expect our method to yield its best performance when applied to documents which have a large number of references, such as Electronic Thesis and Dissertations (ETDs). Therefore, the next version of the BB-ATC system incorporating the enhancements described above will be deployed and evaluated for classification of a large collection of ETD documents archived in a digital library, such as the Networked Digital Library of Thesis and Dissertations (NDLTD) [43].
- As described in Section 4, we used a group of five postgraduate students to manually index the test corpus and used this data as the gold standard to measure the prediction performance of our system. However, for future studies, we plan to also use a second group of professional cataloguers (e.g., librarians) to index the test dataset and measure the predication performance of our system using both gold standards and compare the results.

7. References

[1] Avancini H., Rauber A. and Sebastiani F. Organizing Digital Libraries by Automated Text Categorization. In: Proceedings of the 8th European Conference on Digital Libraries (ECDL 2004); 2004; Bath, UK; 2004. p. 919-931.

[2] Hunter L. and Cohen K. B., Biomedical Language Processing: What's Beyond PubMed?, *Molecular Cell* 2006; 21, 5: 589-594.

[3] Sebastiani F., Machine learning in automated text categorization, *ACM Computing Surveys (CSUR)* 2002; 34, 1: 1-47.

[4] Golub K., Automated subject classification of textual Web pages, based on a controlled vocabulary: Challenges and recommendations, *New Review of Hypermedia and Multimedia* 2006; 12, 1: 11-27.

[5] Yi K., Automated Text Classification Using Library Classification Schemes: Trends, Issues, and Challenges, *International Cataloguing and Bibliographic Control (ICBC)* 2007; 36, 4: 78-82.

[6] Markey K., Forty Years of Classification Online: Final Chapter or Future Unlimited?, *Cataloging & Classification Quarterly* 2006; 42, 3: 1-63.

[7] Dewey M., *Dewey Decimal Classification (DDC)*, (Online Computer Library Center (OCLC), Dublin, Ohio, USA, 1876-2010), http://www.oclc.org/us/en/dewey (accessed February 2011)

[8] Putnam H., *Library of Congress Classification (LCC)*, (Library of Congress, Cataloging Policy and Support Office, Washington, DC, USA, 1897-2010), http://www.loc.gov/catdir/cpso/lcc.html (accessed February 2011)

[9] Godby C. J. and Smith D., *Scorpion*, (OCLC Online Computer Library Center, Inc., 2000-2002), http://www.oclc.org/research/activities/past/orprojects/scorpion/default.htm (accessed February 2011)

[10] *OCLC (Online Computer Library Center*), http://www.oclc.org/ (accessed February 2011)

[11] Larson R. R., Experiments in automatic Library of Congress Classification, *Journal of the American Society for Information Science* 1992; 43, 2: 130-148.

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

[13] Chung Y.-M. and Noh Y.-H., Developing a specialized directory system by automatically classifying Web documents, *Journal of Information Science* 2003; 29, 2: 117-126.

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

[15] Frank E. and Paynter G. W., Predicting Library of Congress classifications from Library of Congress subject headings, *Journal of the American Society for Information Science and Technology* 2004; 55, 3: 214-227.

[16] Wang J., An extensive study on automated Dewey Decimal Classification, *Journal of the American Society for Information Science and Technology* 2009; 60, 11: 2269-2286.

[17] Golub K., Ardö A., Mladenić D. and Grobelnik M. Comparing and Combining Two Approaches to Automated Subject Classification of Text. *Research and Advanced Technology for Digital Libraries*. Springer Berlin / Heidelberg, 2006, p. 467-470.

[18] Sen P., Namata G., Bilgic M., Getoor L., Galligher B. and Eliassi-Rad T., Collective Classification in Network Data, *AI Magazine* 2008; 29, 3.

[19] Joorabchi A. and Mahdi A. E. Leveraging the Legacy of Conventional Libraries for Organizing Digital Libraries. *Research and Advanced Technology for Digital Libraries*. Springer Berlin / Heidelberg, 2009, p. 3-14.
[20] Mahdi A. E. and Joorabchi A., A Citation-based approach to automatic topical indexing of scientific literature, *Journal of Information Science* December 2010; 36, 6: 798-811.
[21] Garfield E., Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas, *Science* 1955; 122, 3159: 108-111.
[22] Aljaber B., Stokes N., Bailey J. and Pei J., Document clustering of scientific texts using citation contexts, *Information Retrieval* 2009; 13, 2: 101-131.
[23] Cao M. D. and Gao X. Combining Contents and Citations for Scientific Document Classification. In: S. Zhang and R. Jarvis, (eds.). *AI 2005: Advances in Artificial Intelligence*. Berlin / Heidelberg: Springer, 2005, p. 143-152.
[24] Bradshaw S. Reference Directed Indexing: Redeeming Relevance for Subject Search in Citation Indexes. In: w. kit and I. T. Sølvberg, (eds.). *Research and AdvancedTechnology for Digital Libraries*. Berlin / Heidelberg: Springer, 2003, p. 499-510.
[25] Bradshaw S. and Hammond K. Automatically indexing documents: content vs. reference. In: 7th international conference on Intelligent user interfaces; 2002; San Francisco, California, USA: ACM; 2002.
[26] Ritchie A., Robertson S. and Teufel S. Comparing citation contexts for information retrieval. In: 17th ACM conference on Information and knowledge management; 2008; Napa Valley, California, USA: ACM; 2008.
[27] Ritchie A., Teufel S. and Robertson S. How to find better index terms through citations. In: Workshop on How Can Computational Linguistics Improve Information Retrieval?; 2006; Sydney, Australia: Association for Computational Linguistics; 2006.
[28] Ritchie A., Teufel S. and Robertson S. Using Terms from Citations for IR: Some First Results. In: C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven and R. W. White, (eds.). *Advances in Information Retrieval*. Berlin / Heidelberg: Springer, 2008, p. 211-221.
[29] Slavic A., Interface to classification: some objectives and options, *Extensions and Corrections to the UDC* 2006; No. 28.
[30] Reiner U. Automatic Analysis of Dewey Decimal Classification Notations. *Data Analysis, Machine Learning and Applications*. 2008, p. 697-704.
[31] Otlet P. and Fontaine H. L., *The Universal Decimal Classification (UDC)*, (UDC Consortium (UDCC), Hague, Netherlands, 1991-2011), http://www.udcc.org/ (accessed February 2011)
[32] Möller G., Carstensen K.-U., Diekmann B. and Wätjen H. Automatic Classification of the World-Wide Web using the Universal Decimal Classification. In: R. Decker and W. Gaul, editors. Proceedings of the 23rd Annual Conference of the German Classification Society (GfKl); 1999; Bielefeld: Springer-Verlag; 1999. p. 231-238.
[33] Giles C. L., Kurt D. B. and Steve L. CiteSeer: an automatic citation indexing system. In: Proceedings of the third ACM conference on Digital libraries; 1998; Pittsburgh, Pennsylvania, United States: ACM; 1998.
[34] Meier W., *eXist-DB*, (exist-db.org, Released under the open source GPL licence, 2009), http://exist.sourceforge.net/ (accessed February 2011)
[35] *Google Books Search (GBS) engine*, (Google, 2004), http://books.google.com/ (accessed February 2011)
[36] *WorldCat*, (Online Computer Library Center (OCLC), Dublin, Ohio, USA, 2001-2010), http://www.oclc.org/worldcat/default.htm (accessed February 2011)
[37] *WorldCat Search API*, (OCLC - WorldCat, 2009), http://worldcat.org/devnet/wiki/SearchAPIDetails (accessed February 2011)
[38] *MARC standards*, (Library of Congress Network Development and MARC Standards Office, 1999), http://www.loc.gov/marc/ (accessed February 2011)
[39] Councill I. G., Giles C. L. and Kan M. Y. ParsCit: An open-source CRF reference string parsing package. In: Language Resources and Evaluation Conference (LREC 08); 2008 May; Marrakesh, Morrocco; 2008.
[40] O'Madadhain J., Fisher D., Nelson T., White S. and Boey Y.-B., *JUNG 2.0*, (Released under the open source GPL licence, 2009), http://jung.sourceforge.net/index.html (accessed February 2011)
[41] de Andrés A., Evaluating research using impact and Hirsch factors, *Europhysics News* 2011; 42, 2: 29-31.
[42] BRIN S., *A Library to Last Forever*, (The New York Times, 8 October 2009), http://www.nytimes.com/2009/10/09/opinion/09brin.html?_r=1 (accessed June 2010)
[43] *Networked Digital Library of Thesis and Dissertations*, (NDLTD, 1996-2010), http://www.ndltd.org (accessed February 2011)

