Genre Classification in Automated Ingest and Appraisal Metadata

Yunhyong Kim and Seamus Ross

Digital Curation Centre (DCC)
&
Humanities Adavanced Technology Information Institute (HATII)
University of Glasgow
Glasgow, UK

Abstract. Metadata creation is a crucial aspect of the ingest of dig- ital materials into digital libraries. Metadata needed to document and manage digital materials are extensive and manual creation of them ex- pensive. The Digital Curation Centre (DCC) has undertaken research to automate this process for some classes of digital material. We have segmented the problem and this paper discusses results in genre clas- sification as a first step toward automating metadata extraction from documents. Here we propose a classification method built on looking at the documents from five directions; as an object exhibiting a specific vi- sual format, as a linear layout of strings with characteristic grammar, as an object with stylo-metric signatures, as an object with intended meaning and purpose, and as an object linked to previously classified objects and other external sources. The results of some experiments in relation to the first two directions are described here; they are meant to be indicative of the promise underlying this multi-facetted approach.

1 Background and Objective

Construction of persistent, cost-contained, manageable and accessible digital collections depends on the automation of appraisal, selection, and ingest of digital material. Descriptive, administrative, and technical metadata play a key role in the management of digital collections ([37],[21]). As DELOS/NSF ([13],[14],[21]) and PREMIS working groups ([34]) noted metadata are expensive to create and maintain. Digital objects are not always accompanied by adequate metadata and the number of digital objects being created and the variety of such objects is increasing at an exponential rate. In response, the manual collection of metadata can not keep pace with the number of digital objects that need to be documented. It seems reasonable to conclude that automatic extraction of metadata would be an invaluable step in the automation of appraisal, selection, and ingest of digital material. ERPANET's ([17]) Packaged Object Ingest Project ([18]) identified only a limited number of automatic extraction tools mostly geared to extract technical metadata (e.g.[29],[31]), illustrating the intensive manual labour required in the ingest of digital material into a repository. Subsequently

substantial work on descriptive metadata extraction has emerged: e.g. extrac- tion from structured documents have been attempted by MetadataExtractor from University of Waterloo ([27]), Dublin Core Metadata Editor ([11]) and Automatic Metadata Generation (AMG) at the Catholic University of Leuven([2]), and the extraction of bibliographic information from medical articles, based on the detection of contiguous blocks and fuzzy pattern matching, is available from Medical Article Record System (MARS) ([42]) developed at the US National Library of Medicine (NLM)([30]). There have also been previous work on metadata extraction from scientific articles in postscript using a knowledge base of stylistic cues ([19],[20]) and, from the language processing community, there have been results in automatic categorisation of emails ([6],[24]), text categorisation ([39]) and document content summarisation ([43]). Other communities have used image analysis for information extraction from the Internet ([3]), document white space analysis ([9]), graphics recognition in PDF files ([41]), and algorithms for page segmentation ([40]). Despite the wealth of research being conducted, no general tool has yet been developed which can be employed to extract meta- data from digital objects of varied types and genres, nor are there dependable extraction tools for the extraction of deeper semantic metadata such as content summary. The research in this paper is motivated by an effort to address this problem by integrating the methods available in the area to create a prototype tool for automatically extracting metadata across many domains at different semantic levels. This would involve:

- constructing a well-structured experimental corpus of one file type (for use in this and future related research);
- summarising and integrating existing research related to automatic metadata extraction;
- determining the limit and scope of metadata that can be extracted and building a prototype descriptive and semantic metadata extraction tool applicable across many domains;
- extending the tool to cover other file types and metadata ; and, – integrating it with other tools to enable automatic ingest, selection and/or appraisal.

The initial prototype is intended to extract Genre, Author, Title, Date, Iden- tifier, Pagination, Size, Language, Keywords, Composition (e.g. existence and proportion of images, text and links) and Content Summary. In the present paper, we discuss genre classification of digital documents represented in PDF ([32]) as a step towards acquiring the appropriate metadata. The term genre does not always carry a clear meaning. We follow the definition of Kessler ([25]) who refers to genre as “any widely recognised class of texts defined by some common communicative purpose or other functional traits, provided the func- tion is connected to some formal cues or commonalities and that the class is extensible”. For instance, a scientific research article is a theoretical argument or communication of results relating to a scientific subject usually published in a journal and often starting with a title, followed by author, abstract, and body

of text,finally ending with a bibliography. One important aspect of genre classification is that it is distinct from subject classification which can coincide over many genres (e.g. a mathematical paper on number theory versus a news article on the proof of Fermat's Last Theorem). The motivation for starting with genre classification is as follows:

- Identifying the genre first will limit the scope of document forms from which to extract other metadata:
  - The search space for further metadata will be reduced; within a sin- gle genre, metadata such as author, keywords, identification numbers or references can be expected to appear in a specific style and region.
  - A lot of independent work exists for extraction of metadata within a specific genre which can be combined with a general genre classifier for metadata extraction over many domains (e.g. the papers listed at the beginning of this section).
  - Resources available for extracting further metadata is different for each genre; for instance, research articles unlike newspaper articles come with a list of reference articles closely related to the original article leading to better subject classification.
- Scoping new genres not apparent in the context of conventional libraries is necessary.
- Different institutional collecting policies might focus on digital materials in different genres. Genre classification will support automating the identification, selection, and acquisition of materials in keeping with local collecting guidelines.

We have opted to consider 60 genres (Table 1). This list is not meant to represent a complete spectrum of possible genres; it is meant to be a starting point from which to determine what is possible.

We have focused our attention on different genres represented in PDF files. By limiting the research to one file type we hoped to put a boundary on the problem space. The choice of PDF as the format stems from the fact that

- PDF is a widely used format. Specifically, PDF is a common format for digital objects ingested into digital libraries including eprint services.
- It is a portable format, distributed over many different platforms. – There are many tools available for conversion to and from other formats.
- It is a versatile format which includes objects of different type (e.g. images, text, links) and different genres (e.g. data structure, fiction, poetry, research article).

In the experiment which follows we worked with a developmental data set collected via the Internet using a random PDF-grabber which

1. selects a random word from a Spell Checker Oriented Word List (from sourceforge.net),
2. searches the Internet using Google for PDF files containing the chosen word,

Table 1. Scope of genres

Groups	Genres
Book	Academic book, Fiction(book), Poetry(book),Other book
Article	Scientific research article, Other research article, Magazine article, News report
Periodicals	Periodicals, Newsletter
Mail	Email, Letter
Thesis	Thesis, Business/Operational report, Technical report, Misc report
List	List,Catalogue
Table	Calendar, Menu, Other table
Proposal	Grant/Project proposal, Legal appeal/proposal/order
Description	Job/Course/Project description, Product/Application description
Minutes	Minutes, Proceedings
Rules	Instruction/Guideline, Regulations
Other	Abstract,Advertisement, Announcement, Appeal/Propaganda, Biography, Chart/Graph,Contract, Drama, Essay, Exam/Worksheet, Fact sheet,Fiction piece, Forms, Forum discussion, Image, Interview, Lecture notes/presentation, Speech transcript, Manual, Memo, Sheet music, Notice, Posters, Programme, Questionnaire, Q & A, Resume/CV, Review, Slides, Poetry piece, Other genre not listed

3. selects a random PDF file from the returned list and places it in a designated folder.

We collected over 4000 documents in this manner. Labelling of this document corpus is still in progress (for genre classification) and is mostly being carried out by one of the authors. Currently 570 are labelled with one of the 60 genres. A significant amount of disagreement is expected in labelling genre even between human labellers; we intend to cross check the labelled data in two ways:

- We will employ others to label the data to determine the level of disagreement between different human labellers; this will enable us to analyse at what level of accuracy the automated system should be expected perform, while also providing us with a gauge to measure the difficulty of labelling individual genres.
- We will gather PDF files which have already been classified into genres as a fresh test data for the classifier; this will also serve as a means of indexing the performance on well-designed classification standards.

Along with the theoretical work of Biber ([7]) on genre structures, there have been a number of studies in automatic genre classification: e.g. Karlgren and Cutting ([23], distinguishing Press, Misc, Non-fiction and Fiction), Kessler et al. ([25], distinguishing Reportage, Fiction, Scitech, Non-fiction, Editorial and Legal; they also attempt to detect the level of readership - which is referred to as Brow - divided into four levels, and make a decision on whether or not

the text is a narrative), Santini ([38], distinguishing Conversation, Interview, Public Debate, Planned Speech, Academic prose, Advert, Biography, Instruction, Popular Lore and Reportage), and, Bagdannov and Worring ([4], fine-grained genre classification using first order random graphs modeled on trade journals and brochures found in the Oc閿 Competitive Business Archive) not to mention a recent MSc. dissertation written by Boese ([8], distinguishing ten genres of web documents). There are also related studies in detecting document logical structures ([1]) and clustering documents ([5]). Previous methods can be divided into groups which look at one or more of the following:

– Document image analysis –
Syntactic feature analysis –
Stylistic feature analysis –
Semantic structure analysis –
Domain knowledge analysis

We would eventually like to build a tool which looks at all of these for the 60 genres mentioned (see Table 1). The experiments in this paper however are limited to looking at the first two aspects of seven genres. Only looking at seven genres out of 60 is a significant cut back, but the fact that none of the studies known to us have combined the first two aspects for genre classification and that very few studies looked at the task in the context of PDF files makes the experiments valuable as a report on the first steps to a general process. This paper is not meant to be a conclusive report, but the preliminary findings of an ongoing project and is meant to show the promise of combining very different classifying methods in identifying the genre of a digital document. It is also meant to emphasise the importance of looking at information extraction across genres; genre-specific information extraction methods usually depend heavily on the structures held in common by the documents in the chosen domain; by looking at differences between genres we can determine the variety of structures one might have to resolve in the construction of a general tool.

2 Classifiers

The experiments described in this paper require the implementation of two classifiers:

Image classifier: this classifier depends on features extracted from the PDF document when handled as an image.

– It uses the module pdftoppm from XPDF to extract the first page of the document as an image then employs Python's Image Library (PIL) ([35], [33]) to extract pixel values. This is then sectioned off into ten regions for an examination of the number of non-white pixels. Each region is rated as level 0, 1, 2, 3 (larger number indicating a higher density of non-white space). The result is statistically modelled using the Maximum Entropy principle. The tool used for the modelling is MaxEnt for C++ developed by Zhang Le ([26]).

Language model classifier: this classifier depends on an N-gram model on the level of words, Part-of-Speech tags and Partial Parsing tags.

- N-gram models look at the possibility of word w(N) coming after a string of words W(1), W(2), ..., w(N-1). A popular model is the case when N=3. This model is usually constructed on the word level. In this research we would eventually like to make use of the model on the level of Part- of-Speech (POS) tags (for instance, tags which denote whether a word is a verb, noun or preposition) or Partial Parsing (PP) tags (e.g. noun phrases, verb phrases or prepositional phrases). Initially we only work with the word-level model. This has been modelled by the BOW toolkit developed by Andrew McCallum ([28]). We used the default Naiive Bayes model without a stoplist.

Although the tools for extracting the image and text of the documents used in these classifiers are specific to PDF files, a comparable representation can be extracted in other formats by substituting these tools with corresponding tools for those formats. In the worst-case scenario the process can be approximated by first converting the format to PDF, then using the the same tools; the wide distribution of PDF ensures the existence of a conversion tool for most common formats.

Using the image of a text document in the classification of the document has several advantages:

- it will be possible to extract some basic information about documents with- out accessing content or violating password protection or copyright;
- more likely to be able to forgo the necessity of substituting language modeling tools when moving between languages, i.e. it maximises the possibility of achieving a language independent tool;
- the classification will not be solely dependent on fussy text processors and language tools (e.g. encoding requirements, problems relating to special characters or line-breaks);
- it can be applied to paper documents digitally imaged (i.e. scanned) for inclusion in digital repositories without heavily relying on accuracy in character recognition.

3 Experiment Design

The experiments in this paper are the first steps towards testing the following hypothesis:

> Hypothesis A: Given a collection of digital documents consisting of several different genres, the set of genres can be partitioned into groups such that the visual characteristics concur and linguistic characteristics dif- fer between documents within a single group, while visual aspects differ between the documents of two distinct groups.

An assumption in the two experiments described here is that PDF documents are one of four categories: Business Report, Minutes, Product/Application Description, Scientific Research Article. This, of course, is a false assumption and limiting the scope in this way changes the meaning of the resulting statistics considerably. However, the contention of this paper is that high level performance on a limited data set combined with a suitable means of accurately narrowing down the candidates to be labelled would achieve the end objective.

Steps for the first experiment

1. take all the PDF documents belonging to the above four genres (70 documents in the current labelled data),
2. randomly select a third of the documents in each genre as training data (27 documents) and the remaining documents as test data (43 documents),
3. train both the image classifier and language model classifier (on the level of words) on the selected training data,
4. examine result.

Steps for the second experiment

1. using the same training and test data as that for the first experiment,
2. allocate the genres to two groups, each group containing two genres: Group I contains business reports and minutes while Group II contains scientific research articles and product descriptions,
3. train the image classifier to differentiate between the two groups and use this to label the test data as documents of Group I or Group II,
4. train two language model classifiers: Classifier I which distinguishes business reports from minutes and Classifier II which labels documents as scientific research articles or product descriptions,
5. take test documents which have been labelled Group I and label them with Classifier I; take test documents which have been labelled Group II and label them with Classifier II,
6. examine result.

The genres to be placed in Group I and Group II were selected by choosing the partition which showed the highest training accuracy for the image classifier.

4 Results

In the evaluation of the results to follow we will use three indices which are considered standard in a classification tasks: accuracy, precision and recall. Let N be the total number of documents in the test data, N_C the number of documents in the test data which are in class C, T the total number of correctly labelled documents in the data independent of the class, T_C the number of true positives

for class C (documents correctly labelled as class C), and F_C the number of false positives for class C (documents labelled incorrectly as class C). Accuracy is defined to be $A = \frac{T}{N}$ while precision and recall for each class C is defined to be $P_C = \frac{T_C}{(T_C+F_C)}$ and $R_C = \frac{T_c}{N_c}$ respectively.

The precision and recall for the first and second experiments are given in Table 2 and Table 3.

Table 2. Result for first small experiment

Overall accuracy (Language model only): 77%

Genres	Prec.(%)	Rec.(%)
Business Report	83	50
Sci. Res. Article	88	80
Minutes	64	100
Product Desc.	90	90

Table 3. Result for second small experiment

Overall accuracy(Image and Language model: 87.5 %

Genres	Prec.(%)	Rec(%)
Business Report	83	50
Sci. Res. Article	75	90
Minutes	71	100
Product Desc.	90	100

Although the performance of the language model classifier given in Table 2 is already surprisingly high, this, to a great extent, depends on the four cate- gories chosen. In fact, when the classifier was expanded to include 40 genres, the classifier performed only at an accuracy of approximately 10%. When a different set was employed which included Periodicals, Thesis, Minutes and In- struction/Guideline, the language model performs at an accuracy of 60.34%. It is clear from the two examples that such a high performance can not be expected for any collection of genres.

The image classifier on Group I(Periodicals) and Group II(Thesis, Minutes, Instruction/Guideline) performs at an accuracy of 91.37%. The combination of the two classifiers have not been tested but even in the worst-case scenario, where we assume that the set of mislabelled documents for the two classifiers have no intersection, the combined classifier would still show an increase in overall accuracy of approximately 10%.

The experiments show an increase in the overall accuracy when the language classifier is combined with the image classifier. To gauge the significance of the increase, a statistically valid significance test would be required. The experiments here however are intended not to be conclusive but indicative of the promise underlying the combined system.

5 Conclusion and Further Research

5.1 Intended Extensions

The experiments show that, although there is a lot of confusion visually and linguistically over all 60 genres, subgroups of the genres exhibit statistically well-behaved characteristics. This encourages the search for groups which are similar or different visually or linguistically to further test Hypothesis A. To extend the scenario in the experiment to all the genres the following steps are suggested.

1. randomly select a third of the documents in each genre as training data and the remaining documents as test data,
2. train the image and language model classifier on the resulting and test over all genres,
3. try to re-group genres so that each group contain genres resulting in a high level of cross labelling in the previous experiment,
4. re-train and test.

5.2 Employment of Further Classifiers

Further improvement can be envisioned by integrating more classifiers into the decision process. For instance consider the following classifiers.

Extended image classifier: In the experiments described in this paper the image classifier looked at only the first page of the document. A variation or extension of this classifier to look at different pages of the document or several pages of the document will be necessary for a complete image analysis. This would however involve several decisions: given that documents have different lengths, the optimal number of pages to be used needs to be determined, and we need to examine the best way to combine the information from different pages (e.g. will several pages be considered to be one image; if not, how will the classification of synchronised pages be statistically combined to give a global classification).

Language model classifier on the level of POS and phrases: This is a N-gram language model built on the part-of-speech tags of the undelying text of the document and also on partial chunks resulting from detection of phrases.

Stylo-metric classifier: This classifier takes its cue from positioning of text and image blocks, font styles, font size, length of the document, average sentence lengths and word lengths. This classifier is expected be useful for both genre classification (by distinguishing linguistically similar Thesis and Scientific Research Article by say the length of the document) and other bibliographic data extraction (by detecting which strings are the Title and Author by font style, size and position).

Semantic classifier: This classifier will combine extraction of keywords, subjective or objective noun phrases (e.g. using [36]). This classifier is expected to play an important role in the summarisation stage if not already in the genre classification stage.

Classifier based on external information: When the source information of the document is available, such features as name of the journal, subject or address of the webpage and anchor texts can be gathered for statistical analysis or rule-based classification.

5.3 Labelling More Data

To make any reasonable conclusions with this study, further data needs to be labelled for fresh experiments and also to make up for the lack of training data. Although 60 genres are in play, only 40 genres had more than 3 items in the set and only 27 genres had greater than or equal to 15 items available.

6 Putting It into Context

Assuming we are able build a reasonable extractor for genre, we will move on to implementing the extraction of author, title, date, identifier, keywords, language, summarisations and other compositional properties within each specific genre. After this has been accomplished, we should augment the tool to handle subject classification and to cover other file types.

Once the basic prototype for automatic semantic metadata extraction is tamed into a reasonable shape, we will pass the protype to other colleagues in the Digital Curation Centre ([10]) to be integrated with other tools (e.g. technical metadata extraction tools) and standardised frameworks (e.g. ingest or preserva- tion model) for the development of a larger scale ingest, selection and appraisal application. Eventually, we should be able at least to semi-automate essential processes in this area.

Acknowledgements

This research is being conducted as part of The Digital Curation Centre's (DCC) [10] research programme. The DCC is supported by a grant from the United Kingdom's Joint Information Systems Committee (JISC) [22] and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC) [16]. The EPSRC grant (GR/T07374/01) provides the support for the research programme. Additional support for this research comes from the *DE-LOS: Network of Excellence on Digital Libraries* (G038-507618) funded under the European Commission's IST 6th Framework Programme [12]. The authors would like to thank their DCC colleague Adam Rusbridge whose work on ER- PANET's Packaged Object Ingest Project [18] provided a starting point for the current project on automated metadata extraction. We are grateful to the anonymous ECDL reviewers of this paper who provided us with very helpful comments, which enabled us to improve the paper.

Note on website citations: All citations of websites were validated on 29 May 2006.

References

1. Aiello, M., Monz, C., Todoran, L., Worring, M.: Document Understanding for a Broad Class of Documents. International Journal on Document Analysis and Recognition **5(1)** (2002) 1–16.
2. Automatic Metadata Generation: http://www.cs.kuleuven.ac.be/□mdb/amg /documentation.php
3. Arens,A., Blaesius, K. H.: Domain oriented information extraction from the In- ternet. Proceedings of SPIE Document Recognition and Retrieval 2003 **Vol 5010** (2003) 286.
4. Bagdanov, A. D., Worring, M.: Fine-Grained Document Genre Classification Using First Order Random Graphs. Proceedings of International Conference on Docu- ment Analysis and Recognition 2001 (2001) 79.
5. Barbu, E., Heroux, P., Adam, S., Trupin, E.: Clustering Document Images Using a Bag of Symbols Representation. International Conference on Document Analysis and Recognition, (2005) 1216–1220.
6. Bekkerman, R., McCallum, A., Huang, G.: Automatic Categorization of Email into Folders. Benchmark Experiments on Enron and SRI Corpora', CIIR Technical Report, **IR-418** (2004).
7. Biber, D.: Dimensions of Register Variation:a Cross-Linguistic Comparison. Cambridge University Press (1995).
8. Boese, E. S.: Stereotyping the web: genre classification of web documents. Master's thesis, Colorado State University (2005).
9. Breuel, T. M.: An Algorithm for Finding Maximal Whitespace Rectangles at Ar- bitrary Orientations for Document Layout Analysis. 7th International Conference for Document Analysis and Recognition (ICDAR), 66–70 (2003).
10. Digital Curation Centre: http://www.dcc.ac.uk
11. DC-dot, Dublin Core metadata editor: http://www.ukoln.ac.uk/metadata/dcdot/
12. DELOS Network of Excellence on Digital Libraries: http://www.delos.info/
13. NSF International Projects: http://www.dli2.nsf.gov/ intl.html
14. DELOS/NSF Working Groups: Reference Models for Digital Libraries: Ac- tors and Roles (2003) http://www.dli2.nsf.gov /internationalprojects/ work- ing group reports/ actors final report.html
15. Dublin Core Initiative: http://dublincore.org/tools/#automaticextraction
16. Engineering and Physical Sciences Research Council: http://www.epsrc.ac.uk/
17. Electronic Resources Preservation Access Network (ERPANET): http:// www.erpanet.org
18. ERPANET: Packaged Object Ingest Project. http://www.erpanet.org/events/ 2003/rome/presentations/ ross rusbridge pres.pdf
19. Giuffrida, G., Shek, E. Yang, J.: Knowledge-based Metadata Extraction from PostScript File. Proc. 5th ACM Intl. conf. Digital Libraries (2000) 77–84.
20. Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E. A.: Automatic Docu- ment Metadata Extraction using Support Vector Machines. Proc. 3rd ACM/IEEE- CS conf. Digital libraries (2000) 37–48.
21. Hedstrom, M., Ross, S., Ashley, K., Christensen-Dalsgaard, B., Duff, W., Gladney, H., Huc, C., Kenney, A. R., Moore, R., Neuhold, E.: Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation. Report of the European Union DELOS and US National Science Foundation Workgroup on Digital Preservation and Archiv- ing (2003) http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf.

22. Joint Information Systems Committee: http://www.jisc.ac.uk/
23. Karlgren, J. and Cutting, D.: Recognizing Text Genres with Simple Metric using Discriminant Analysis. Proc. 15th conf. Comp. Ling. **Vol 2** (1994) 1071–1075.
24. Ke, S. W., Bowerman, C. Oakes, M. PERC: A Personal Email Classifier. Proceed- ings of 28th European Conference on Information Retrieval (ECIR 2006) 460–463.
25. Kessler, B., Nunberg, G., Schuetze, H.: Automatic Detection of Text Genre. Proc. 35th Ann. Meeting ACL (1997) 32–38.
26. Zhang Le: Maximum Entropy Toolkit for Python and C++. LGPL license, http://homepages.inf.ed.ac.uk/s0450736/maxent toolkit.html
27. MetadataExtractor: http://pami-xeon.uwaterloo.ca/TextMiner/ MetadataExtrac- tor.aspx
28. McCallum, A.: Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering. (1998) http://www.cs.cmu.edu/ mccallum/bow/
29. National Archives UK: DROID (Digital Object Identification). http: //www. na- tionalarchives. gov.uk/ aboutapps/pronom/droid.htm
30. Natinal Library of Medicine US: http://www.nlm.nih.gov/
31. National Library of New Zealand: Metadata Extraction Tool. http://www. natlib. govt.nz/en/whatsnew/4initiatives.html#extraction
32. Adobe Acrobat PDF specification: http://partners.adobe.com/ public/developer/ pdf/index reference.html
33. Python Imaging Library: http://www.pythonware.com/products/pil/
34. PREMIS (PREservation Metadata: Implementation Strategy) Working Group: http://www.oclc.org/research/projects/pmwg/
35. Python: http://www.python.org
36. Riloff, E., Wiebe, J., and Wilson, T.: Learning Subjective Nouns using Extraction Pattern Bootstrapping. Proc. 7th CoNLL, (2003) 25–32.
37. Ross S and Hedstrom M.: Preservation Research and Sustainable Digital Libraries. International Journal of Digital Libraries (Springer) (2005) DOI: 10.1007/s00799- 004- 0099-3.
38. Santini, M.: A Shallow Approach To Syntactic Feature Extraction For Genre Clas- sification. Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 04) (2004).
39. Sebastiani F.: 'Machine Learning in Automated Text Categorization', ACM Com- puting Surveys, **Vol. 34** (2002) 1-47
40. Faisal Shafait, Daniel Keysers, Thomas M. Breuel, "Performance Comparison of Six Algorithms for Page Segmentation", 7th IAPR Workshop on Document Analysis Systems (DAS) (2006).368–379.
41. M. Shao, M. and Futrelle, R.: Graphics Recognition in PDF document. Sixth IAPR International Workshop on Graphics Recognition (GREC2005), 218–227.
42. Thoma,G.: Automating the production of bibliographic records. R&D report of the Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine, 2001.
43. Witte, R., Krestel, R. and Bergler, S.: ERSS 2005:Coreference-based Summariza- tion Reloaded. DUC 2005 Document Understanding Workshop, Canada