# Automatic extraction of metadata from scientific publications for CRIS systems

Aleksandar Kovačević, Dragan Ivanović, Branko Milosavljević, Zora Konjović and Dušan Surla

**Authors:** Aleksandar Kovačević, Faculty of Technical Sciences, University of Novi Sad, Serbia. Email: kocha78@uns.ac.rs

Dragan Ivanović, Faculty of Technical Sciences, University of Novi Sad, Serbia. Email: chenejac@uns.ac.rs

Branko Milosavljević, Faculty of Technical Sciences, University of Novi Sad, Serbia. Email: mbranko@uns.ac.rs

Zora Konjović, Faculty of Technical Sciences, University of Novi Sad, Serbia. Email: ftn\_zora@uns.ac.rs

Dušan Surla, Faculty of Science, University of Novi Sad, Serbia. Email: surla@uns.ac.rs

**Abstract**

**Purpose –** The aim of the research is to develop a system for automatic extraction of metadata from scientific papers in PDF format for the information system for monitoring the scientific research activity of the University of Novi Sad (CRIS UNS). By this, inputting metadata into CRIS UNS is reduced to the correction or update of the extracted metadata. This system will also help in the process of synchronising metadata from CRIS UNS with other institutional repositories.

**Design / methodology / approach –** The system is based on machine learning and performs automatic extraction and classification of metadata in eight pre-defined categories: *Title*, *Authors*, *Affiliation*, *Address*, *Email*, *Abstract*, *Keywords* and *Publication Note*. The extraction task is realised as a classification process. For the purpose of classification each row of text is represented with a vector that comprises different features: formatting, position, characteristics related to the words, etc. Experiments were performed with standard classification models: *Decision Tree*, *Naive Bayes*, *K-nearest Neighbours* and *Support Vector Machines*. Both a single classifier with all eight categories and eight individual classifiers were tested. Classifiers were evaluated using the five-fold cross validation, on a manually annotated corpus comprising 100 scientific papers in PDF format, collected from various conferences, journals and author's personal web pages.

**Findings –** Based on the performances obtained on classification experiments, eight separate *Support Vector Machines* (SVM) models (each of which recognises its corresponding category) were chosen. All of these eight models were established to have a good peformance. The F-measure was over 85% for almost all of the classifiers and over 90% for the most of them.

**Research limitations / implications –** Automatically extracted metadata can not be directly entered into CRIS UNS but require control of the curators.

**Practical implications –** The proposed system for automatic metadata extraction using *Support Vector Machines* model was integrated into the software system, CRIS UNS. Metadata extraction has been tested on the publications of researchers from the Department of Mathematics and Informatics of the Faculty of Sciences in Novi Sad. Analysis of extracted metadata from these publications showed that the performances of the system for the previously unseen data are in accordance with those obtained by the cross validation from eight separate SVM classifiers.

**Originality / value –** A fully automated system for metadata extraction from sientific papers was developed. The system is based on the SVM classifier and open source tools, and is capable of extracting eight types of metadata from scientific articles of any format that can be converted to PDF. Although developed as part of CRIS UNS, the proposed system can be integrated into other CRIS systems, as well as institutional repositories and library management systems.

**Article type**: Research paper

**Keywords**: automatic metadata extraction, CRIS, classification, Support Vector Machines, PDF format

## Introduction

Metadata as a term is used to denote all of the information that describes the characteristics of objects that are stored in digital repositories. The need for quality metadata has become greater with the development of the Internet, which is becoming a powerful resource for the dissemination and exchange of information. Metadata are a fundamental component of digital libraries as well as initiatives such as the *Semantic Web* (<http://www.w3.org/2001/sw/>) and *Open Archives Initiative* (<http://www.openarchives.org/>). Institutional repositories such as *DSpace* (<http://www.dspace.org/>), *Fedora* (<http://fedora-commons.org/>) and *EPrints* (<http://www.eprints.org/>) use metadata to enable efficient access and retrieval of digital content. An enormous amount of digital resources requiring metadata represent a major challenge for these initiatives, libraries and repositories For example, a project which would aim at forming a digital repository of all publicly available scientific papers in the UK, would have to face a collection with the growing pace of 100,000 papers per year that require manual metadata entry (Adams 2009). Ideally, the metadata should be entered by the authors, but authors rarely do that even when they are provided with the appropriate tools (Crystal and Land 2003). According to Chrystal it would take about 60 employee-years to create metadata for 1 million documents (Crystal and Land 2003). Based on overwhelming cost of manually entering metadata, it is obvious that there is a need for tools that enable their automatic extraction. The U.S. Library of Congress, i.e. its Directorate for cataloguing, has recognised this problem (Adams 2009) and sponsored the *Automatic Metadata Generation Applications* (AMEGA) project (Greenberg et al., 2006).

Based on the development of automatic indexing of digital content and the fact that it is less costly than manual indexing (Anderson and Pérez-Carballo 2001), it can be assumed that, over time, automated metadata extraction is to become more efficient, cheaper and more consistent. Although previous studies show that automatic generation of metadata provides acceptable performance (Liddy et al., 2002; Han et al., 2003; Peng and Mccallum 2004; Takasu 2003), researchers generally conclude that the best results are achieved by integrating automated and manual methods (Schwartz 2001).

This paper presents a method for automatic extraction of metadata from scientific articles in PDF format which is designed as an integral part of the information system for monitoring the scientific research activity - CRIS UNS (<http://cris.uns.ac.rs/>). The method is implemented as a complement to manual metadata entry in the sense that the extraction results are offered to the curator to inspect and correct before storing them into the repository.

### Motivation

The basic motivation for the development of the proposed system is to provide an additional way of entering metadata into CRIS UNS. The process of manual metadata entry is time consuming and error-prone. By using automatic extraction of metadata this process is reduced to the correction or update of the extracted metadata.

During large scientific projects and studies, a substantial amount of scientific literature is collected. For the purposes of efficient search, retrieval and report generation, metadata for these publications is stored in CRIS UNS. Automatically generated metadata, despite its lesser quality than the manually entered one, is a good alternative for dramatically speeding up this kind of tasks.

The integration of CRIS UNS with existing institutional repositories is being performed. During the integration process there is a need for metadata synchronization. When this type of work is done, the cases in which the two systems contain different metadata relating to the same digital document are frequent. Metadata from the digital document extracted by the system presented in this paper can be of great importance for dealing with such situations.

### Metadata extraction

Metadata extraction system presented in this paper performs fully automated extraction from scientific articles in PDF format. The decision to process papers in PDF format is based on the fact that most of the scientific publications are available in this format. Regardless of that fact, the system can be applied to papers of any format that can be converted to PDF for example: Microsoft Word, Latex, PostScript, etc.

The system is based on machine learning (ML) methods, i.e., classification. Metadata is classified in eight pre-defined categories: *Title*, *Authors*, *Affiliation*, *Address*, *Email*, *Abstract*, *Keywords* and *Publication Note*. Experiments were performed with the standard classification models: *Decision Tree* (DT), *Naive Bayes* (NB), *K-nearest Neighbours* (KNN) and *Support Vector Machines* (SVM).

The experiments were conducted with a single classifier with all eight categories and eight individual classifiers constructed by decomposing a multi-category classification problem into multiple binary problems. The classifiers were evaluated using the five-fold cross-validation on a manually annotated corpus of scientific papers in PDF format, collected from different conferences, journals and author's personal Web pages. As the measure of performance, standard measures for these types of tasks were used, such as precision, recall and F-measure.

The best results were achieved by eight separate *Support Vector Machines* models. Based on that fact, they were chosen as a classification model for the proposed system. The overall performances were good, F-measures was over 85% in almost all categories: *Title* (F-measure of 98.77%), *Email* (F-measure of 98.41%), *Authors* (F-measure of 92.13%), *Abstract* (F-measure of 91.52%), *Affiliation* (F-measure of 90.37%), *Address* (F-measure of 87.80%) and *Publication Note* (F-measure of 86.83%). The value of the F-measure for the *Keywords* category was somewhat lower – 81.42%.

The rest of the paper is organised as follows. Section 2 reviews the related works. Current Research Information System CRIS UNS is described in Section 3. Section 4 presents the system for automatic metadata extraction. The corpus used for evaluation is described in Section 5. The experimental results are given in Section 6. Integration with CRIS UNS and verification of the proposed system are given in Section 7. Section 8 concludes the paper.

## Related work

Automatic extraction of metadata is a research discipline that is constantly gaining in importance. Methodologies in this field can generally be classified into two categories: rule-based approaches and approaches based on machine learning. The following are the major approaches.

### Rule-based systems

Rule-based systems use a set of predefined instructions that specify how to extract metadata from documents (Bergmark 2000; Klink et al., 2000; Mao et al., 2004). Liddy and colleagues have developed a rule-based system that employs natural language processing technologies in order to extract metadata from educational materials (Liddy et al., 2002; Yilmazel et al., 2004). Most of the systems in this category, when writing rules, exploit the fact that scientific articles usually follow a common format. The approach presented in (Giuffrida et al., 2000) uses the rules of the form "title is usually found in the first parts of the text and has the largest font" in order to identify the metadata in articles that are in Postscript format. Rules based on formatting were also used in (Mao et al., 2004) to perform automatic metadata extraction from medical journal research papers. Flynn and colleagues use string matching and a set of rules based on the positional relationships for the extraction of metadata from heterogeneous collections of scientific papers and project reports (Flynn et al., 2007). Formatting style and font information are combined in (Groza et al., 2009) in order to create algorithms for the extraction of: title, author, sections and references from scientific articles in PDF format. The approach presented in (Ojokoh 2009) combines segmentation based on keywords and regular expressions in order to extract general metadata from documents such as title, table of contents, abstract, preface etc. The authors focus mainly on theses and dissertations in PDF format. In recent paper (Abugessaisa 2010) the author presents the results of conformance test on the metadata held at the *Swedish Land Survey* (SLS) and EU INSPIRE directive (<http://inspire.jrc.ec.europa.eu/>) implementation rules (IRs) for metadata. The results reveal a gap between the existing metadata at the SLS and INSPIRE directive IRs. In order to bridge this gap an automatic method for the extraction of geospatial metadata by applying techniques from ontology engineering is introduced. The metadata is extracted from product documents which are manually created rich text files that describe geospatial datasets. Devignes and colleagues (Devignes et al., 2010) present a system for automatic metadata extraction with a goal of creating a repository of relevant biological resources. The metadata is obtained by automatically parsing entries from *Molecular Biology Database* an existing catalogue of biological databases.

Rule based systems tend to have very good performance. However, due to the strong connection between the rules and the layout or text of the scientific articles, such systems are less adaptable and robust than those based on machine learning. Also rule construction is a time-consuming task (Klink et al., 2000), and the complexity of the system rapidly increases with the number of rules.

### Systems based on machine learning

The systems in this category usually formalise the problem of automatic metadata extraction as a classification or a sequence tagging task. Hu and colleagues (Hu et al., 2006) used machine learning methods for the extraction of titles from a wide range of documents. The authors in (Han et al., 2003) realised the task of metadata recognition as the classification of rows in the document. The classification is performed by *Support Vector Machines*, and each row is represented by its linguistic characteristics. The authors achieved a very good performance on a collection of scientific articles in the terms of accuracy, precision and recall. Peng and McCallum also carried out research on the extraction of metadata from scientific papers (Peng and Mccallum 2004). They use a sequence tagging technique called *Conditional Random Fields* (CRF) and different types of word specific and layout features, along with an external lexicon, in order to tag each of the words in the header of the article with the appropriate metadata category. Cui used *Hidden Markov models* (HMM) along with information about the layout to identify the metadata in scientific papers (Cui 2009). In a more recent paper the same author used HMM along with text block visual features to recognize six types of metadata (Cui 2010). The presented system vas evaluated on dataset consisting of papers from VLDB (Very Large Databases) conferences. System presented in (Marinai 2009) combines low level document image processingand layout analysis with the neural network classifier in order to recognise titles and authors from PDF papers in the field of Computer Science. As a way of boosting performance the authors use a well know database of Computer Science publications called *Digital Bibliography & Library Project* (DBLP) (<http://www.informatik.uni-trier.de/~ley/db/>). In recent paper (Lin et al., 2010) automatically extract metadata from clinical research articles. These metadata fields include metadata about the authors, as well as information concerning the parameters and medical intervention methods used in a clinical study.

Machine learning methods usually perform well on relatively homogeneous collections such as research papers. Disadvantages of these methods are that they depend on manual annotation of training sets which requires a lot of time and effort. Performance of these methods also tends to decline when applied to the highly heterogeneous collections i.e. a classifier trained on a corpus of scientific papers would probably perform poorly in extracting metadata from PhD thesis and vice versa.

Since our goal is to automatically extract metadata from scientific papers published in proceedings and journals of various layouts we decided to use machine learning approach as manually constructing rules in this case would be a very time consuming process. Most of the ML approaches use only a few types of features e.g. layout (Cui 2009; Marinai 2009) or linguistic (Han et al., 2003). With the goal of boosting performance we decided to use five types of features in our system (see section 4.1). Instead of using lexicons or databases with the names of researches, institutions etc. (Peng and Mccallum 2004; Marinai 2009) we used Named Entity Recognition (see section 4.1) in order to be able to extract researches and organizations that are not contained in lexicons (databases). As our goal is to process papers in PDF format and to use formatting features in contrast to most of the ML approaches we opted not to use an existing dataset (Han et al., 2003; Peng and Mccallum 2004; Cui 2009) which consists of textual content stripped of formatting information such as font size, colour, style etc. We also differ from other systems that use PDF as input (Marinai 2009; Mao et al., 2004; Groza et al., 2009; Cui 2010) in a sense that our dataset includes a wide variety of proceedings and journals with varying layouts.

## Current Research Information System of the University of Novi Sad - CRIS UNS

Research management systems take significant role in the development of science (Zimmerman 2002). Those systems collect, preserve and disseminate data about institutions, researchers, research projects, equipment, published results and other relevant data for scientific-research activity. The European Union encourages the development of national research management systems in accordance with the CERIF standard (*the Common European Research Information Format* - <http://www.eurocris.org/cerif/introduction/>), because its goal is to achieve maximum competitiveness of Europe at all levels of research activity. CERIF compatible research management systems are called CRIS (*Current Research Information System*). CERIF provides a data model that enables interoperability between research management systems (Asserson et al., 2002). CERIF defines a physical data model (Jörg et al., 2009a) and an XML data exchange format (Jörg et al., 2009b).

CRIS UNS is a CERIF compatible research management system which has been being developed since 2008 at the University of Novi Sad in the Republic of Serbia. Experience gained from developing of the BISIS library information system (<http://www.bisis.uns.ac.rs/>) has been used for developing CRIS UNS. The BISIS system has been being developed since 1993 at the University of Novi Sad. The current version 4 is based on XML technology. Within the version, an XML editor for cataloguing in the UNIMARC and MARC21 format (Dimić and Surla 2009; Dimić et al., 2010) that uses the component for generating catalogue cards (Rađenović et al., 2009) and text server (Milosavljević et al., 2010) has been developed. The paper (Boberić and Surla 2009) describes an XML editor for searching and downloading bibliographic records in accordance with Z39.50 standard, and the papers (Tešendić et al., 2009; Milosavljević and Tešendić 2010) describe a library circulation system. Distributed library catalogues in BISIS are discussed in the paper (Vidaković et al., 2009). The papers (Belić and Surla 2008a; Belić and Surla 2008b) describe modelling and implementation of web-based system for entering data about published scientific-research outputs in the UNIMARC format.

The first has coverd development of a system for entering metadata about published scientific-research outputs in the following forms: papers published in journals, papers published on scientific conferences, monographs, papers published in monographs. The system is built on the CERIF compatible data model based on the MARC21 format which is presented in the paper (Ivanović et al., 2011b). The system implementation is described in papers (Ivanović et al., 2010; Milosavljević et al., 2011). Published results from the system are available to anonymous user via the Internet. The system is implemented as web application that enables authors to input metadata about their own research, without the knowledge of the CERIF standard and the MARC21 format. Moreover, the system is in accordance with the CERIF standard and meets requirements prescribed by Ministry of Science and Technological Development of the Republic of Serbia in the field of scientific-research results evaluation. Therefore, the system data model is extended with necessary entities (Ivanović et al., 2011a). The system data model and architecture allow easy integration of the system with library information systems and interoperability with other CERIF compatible national CRIS systems of European countries.

## Automatic metadata extraction system

The aim of the system presented in this paper is to automatically extract and categorise metadata in eight pre-defined categories: *Title*, *Authors*, *Affiliation*, *Address*, *Email*, *Abstract*, *Keywords* and *Publication Note*. The task of automatic metadata extraction can be viewed as an *information extraction* (IE) problem. The work by Han et al. and Chieu (Chieu and Ng 2002; Han et al., 2003) suggests that IE tasks can be successfully solved by classification. Based on that, in the proposed system, the task of automatic metadata extraction is realised as a classification process.

### Feature extraction

The system performs the extraction only from the first page of the scientific article since that page, in majority of the cases, contains all available metadata. In accordance to the previous research one row of text in the paper is treated as a classification unit (Han et al., 2003).

As a first step in the extraction process the conversion from PDF into HTML format is performed. The choice to convert PDF format into HTML was made because this format preserves the formatting information during conversion. This information is crucial for the recognition of metadata, since the classification method presented in this paper exploits characteristics such as style, font size, layout etc.

After the conversion process, feature extraction is performed for each text row on the first page (Figure 1). Most of the previous approaches in the field of information extraction use word-specific features (Ramshaw and Marcus 1995; Lawrence et al., 2002; Takeuchi and Collier 2002). Line-specific features are also suggested to be useful by recent research (McCallum et al., 2000; Han et al., 2003). In the classification method presented in this paper both word and line specific features are combined along with other features in order to form a feature vector which represents one row of text. The extracted features can be categorised in the following five categories (Tables 1-4):

* **Formatting features**.This group consists of features obtained from the formatting information of the first page, as returned by the *pdf2html* tool. Examples are: the size and name of the font, presence of bold or italic formatting, etc. (Table 1).
* **Word-specific features**. Decision to use these features is based on the previous research, which showed that certain characteristics of the words such as orthographic have a major impact on the recognition of metadata (Peng and Mccallum 2004). Since the classification model presented in this paper uses rows as a classification unit, all of the words in a row are considered in the extraction process. Examples are: "does the row contains an abbreviation", "does the row contains a number", "does the row contains initials", etc. (Table 2).
* **Named entity recognition features**. Named Entity Recognition (NER) is a well-developed method for automatically placing the words in the text into certain categories (Nadeau and Sekine 2007), such as a person, location, etc. Since author's names and surnames can be recognised as a person entity and addresses and affiliations usually contain names of locations and institutions, it is reasonable to conclude that including NER entity types in the feature set would benefit the classification model. Entity types were extracted using the NER tool developed at Stanford University (Finkel et al., 2005). This tool has been selected due to its good performance and the fact that it is frequently used in information extraction tasks. The following types of entities are identified: persons, locations and organisations (Table 3).
* **Layout features**.Layout features include such characteristics of the text row as: absolute and relative position, distance (number of rows) from some particular rows, e.g., the row that contains the word "abstract" or "keywords" (Table 4).
* **Lexicon features**. The fact that particular types of metadata in scientific papers contain specific words or types of words is exploited when extracting the features from this category. For example the affiliations of authors often contain keywords such as "university" or "faculty", while the notes about publication type contain terms such as "proceedings" or "journal" and so on. Hand-crafted lexicons were constructed in order to extract these types of features (Table 4).

|  |  |
| --- | --- |
| **Formatting features** | |
| isMaxFont | row has the largest font size |
| isMinFont | row has the smallest font size |
| absPosLeft | the absolute position of the row relative to the left edge of the document as returned by the pdf2html tool |
| absPosTop | the absolute position of the row relative to the top edge of the document as returned by the pdf2html tool |
| styleName | css style name (internal for pdf2html). Useful for distinguishing rows with different formatting style |
| styleFontFamliy | the name of the font used in the row |
| fontSize | the size of the font used in the row |
| lineHeight | the height of the row (as returned by the pdf2html) |
| styleColor | the colour of the text in the row |
| ahrefTags | the number of HTML links in the row |
| boldTags | the number of words with the bold style font |
| italicTags | the number of words with the italic style font |

Table 1. Formatting features.

|  |  |
| --- | --- |
| **Word-specific features** | |
| wordCount | the number of words in the row |
| containsAcro | row contains at least one abbreviation |
| containsAddress | row contains at least one street address |
| containsAllCaps | row contains at least one word with all capital letters |
| containsAllDigits | row contains at least one word with all digits |
| containsCapLetter | row contains at least one word which is just one capital letter |
| containsDash | row contains at least one word which contains a dash character |
| containsDigits | row contains at least one word which contains at least one digit |
| containsDot | row contains at least one word which contains a dot character |
| containsEmail | row contains at least one email address |
| containsPartialEmail | row contains at least one partial email e.g. has just "@uns.ac.rs". Useful for recognising the cases where two or more authors have an email address on the same Internet domain e.g. "{author1, author2} @uns.ac.rs " |
| containsInitCap | row contains at least one word that starts with a capital letter |
| containsLonlInit | row contains at least one word that is an initial e.g. "A." |
| containsPhOrZip | row contains at least one word that is phone number or a zip code |
| containsPunct | row contains at least one punctuation mark |
| containsSingleChar | row contains at least one word that is just one character |
| containsUrl | row contains at least one URL |
| allInitCaps | all words in the row start with a capital letter |

Table 2. Word-specific features.

|  |  |
| --- | --- |
| **Named entity recognition features** | |
| personCount | the number of person entities recognised in the row |
| locationCount | the number of location entities recognised in the row |
| organisationCount | the number of organisation (institution) entities recognised in the row |

Table 3. Named entity recognition features.

|  |  |  |  |
| --- | --- | --- | --- |
| **Layout features** | | **Lexicon features** | |
| absLoc | ordinal number of the row starting from the top of the page | justKeywords | row contains just the word "keywords" or "keywords:" |
| relativeLoc | The overall number of the rows on the page is divided into 10 parts by ordinal numbers. The relative location of the row represents the part that row belongs to. | keywordsFirst | the word "keywords" or "keywords:" is the first word in the row |
| isFirstRow | row is the first row starting from the top of the page | keywordDict | row contains a term from the *Keyword* metadata lexicon e.g. "general terms", "category", "subject" etc. |
| isSecondRow | The row is the second row starting from the top of the page. Useful because many titles consist of two rows. | justAbstract | row contains just the word "abstract" |
| distFromKeyword | The number of rows between this row and the row that contains just the word "keywords" or -1 if there is no such a word on the page. Useful because the keyword metadata can be spread across more than one row. | affiliationDict | row contains a term from the *Affiliation* metadata lexicon e.g. "university", "faculty", "department" etc. |
| distFromClosestEmail | the number of rows between this row and the closest row that contains an email address | month | row contains a full or abbreviated month name e.g. "January", "Feb." etc. |
| distFromClosestPartialEmail | the number of rows between this row and the closest row that contains a partial Email address | year | row contains a calendar year |
| distFromMaxFontInFirstPart | The number of rows between this row and the row that belongs to the first part (calculated for determining the relative location) and has the largest font in that part. This feature can be considered as the distance from the paper title. | publicationDict | row contains a term from the *Publication note* metadata lexicon e.g. "proceedings", "journal", "vol." etc. |

Table 4. Layout and lexicon features.

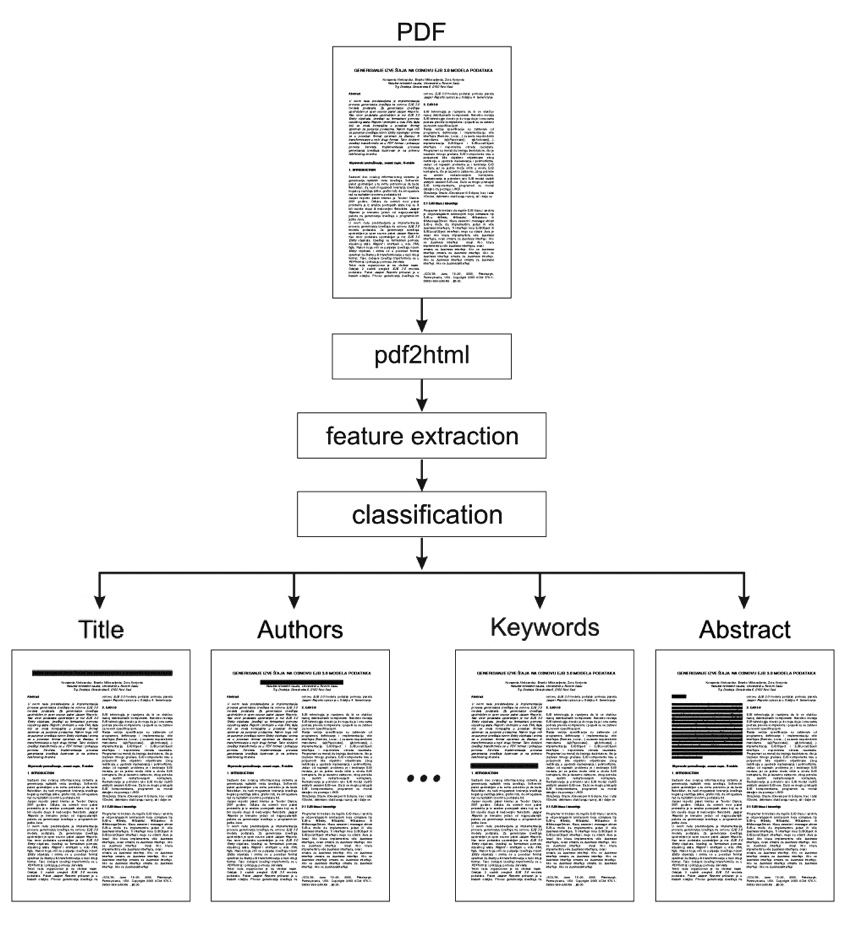


Figure 1. The metadata extraction process.

### Classification

Experiments were conducted with standard classification models: *Decision Tree*, *Naive Bayes*, *K-nearest Neighbours* and *Support Vector Machines*.

**The *Decision Tree*** is a tree whose internal nodes are marked with features, and in which the branches that link those nodes are marked by the values of these features, while the leaves contain class labels (Quinlan 1986; Quinlan 1987).Classification is done by moving down the tree along the path dictated by the feature values until a leaf is reached. This classifier has been successfully applied in IE tasks such as: classification of personal names (Sekine and Grishman 1998), extraction of information from the Web (Hong-ye 2009), sentence analysis (Cardie 1993), etc.

***Naive Bayes*** model belongs to probabilistic classifiers that view the problem of categorisation from the standpoint of conditional probabilities that are determined by the Bayes theorem (Domingos and Pazzani 1997). Heidorn and Wei use NB classifier for extracting metadata from museum specimens labels (Heidorn and Wei 2008). Mcallum and colleagues present an extension to the NB model for text classification used in the automatic construction of Internet portals (McCallum et al., 2000).

***K-nearest Neighbours*** belongs to the category of lazy classification methods. These methods do not create explicit classification models, but rather perform the categorisation by calculating the similarity of the new instance with instances already in training set. The class label of the new instance is determined by the labels of *k* most similar instances. Good examples of use of the *K-nearest Neighbours* for text classification are given in (Larkey and Croft 1996; Y. Yang 2001). The approach described in (Lattner and Herzog 2003) uses the KNN model to classify text in the process of automatic metadata generation from the digital repositories of companies.

***Support Vector Machines*** are a classification model with good generalisation capabilities and ability to handle high dimension data. This technique has been developed to solve binary classification problems. SVM solves the classification problem by finding the maximum margin hyperplane that separates the feature vectors by the class label (Vapnik 1998). For the datasets with a linearly inseparable feature space kernel functions are used (Vapnik 1998). Most commonly used ones are: linear, polynomial and radial basis function, although any function that satisfies Mercer's condition can be used (Vapnik 1998; Cristianini 2000). This technique was successfully applied to the metadata extraction tasks in (Han et al., 2003; Han et al., 2005).

Experiments were performed both with a single classifier with all eight categories and eight individual classifiers. The Performances were calculated by using the five fold cross-validation on manually annotated corpus (see sections 5 and 6).

## The gold standard corpus

In order to form the classification models and evaluate the system presented in this paper, a corpus of 100 scientific articles in the field of Computer Science, i.e. the sub field *Automatic Term Recognition* (ATR) was collected. This corpus is used in a project that deals with the extraction of methodologies from scientific publications (<http://www.informatika.ftn.uns.ac.rs/AleksandarKovacevic/MethodologyExtraction/>). The results of that project, along with the system presented in this paper, are integrated into CRIS UNS as a means of semantically enriching the indexing of scientific articles. The corpus consists of the full text articles from the field of ATR, including papers from *Annual Meeting of the Association for Computational Linguistic* (ACL), *Association for Computing Machinery* (ACM), *Conference on Computational Linguistics* (COLING) and various other conferences, journals and author's personal web pages. Even thou the corpus belongs to a particular sub field of Computer Science, the publications in it come from a large number of different sources which makes it challenging (heterogeneous) enough for the task of automatic metadata extraction. The first pages of the articles in the corpus were manually annotated in eight pre-defined categories: *Title*, *Authors*, *Affiliation*, *Address*, *Email*, *Abstract*, *Keywords* and *Publication Note*. If the row could not be classified into any of the aforementioned categories (the sentences in the "Introduction" section and the like) it was annotated as a separate category called *Other*. The above categories are a part of a standard set of categories (McCallum et al., 2000; Han et al., 2003; Peng and Mccallum 2004; Cui 2009), adapted to the task performed by the system presented in this paper. The corpus size is in accordance with the similar approaches in which the systems are evaluated on a manually annotated corpus (Giuffrida et al., 2000; Yin et al., 2005; Marinai 2009). Annotation was performed using the tool Calisto (<http://callisto.mitre.org/>). Corpus Statistics are given in Table 5.

|  |  |  |
| --- | --- | --- |
| ***Category*** | ***Number of rows*** | ***Percentage*** |
| *Title* | 159 | 4,12% |
| *Authors* | 233 | 6,05% |
| *Affiliation* | 315 | 8,18% |
| *Address* | 184 | 4,77% |
| *Email* | 180 | 4,67% |
| *Abstract* | 515 | 13,37% |
| *Keywords* | 38 | 0,98% |
| *Publication Note* | 92 | 2,38% |
| *Other* | 2133 | 55,40% |
| **Overall:** | 3850 |  |

Table 5. The statistic of the gold standard corpus.

## Experimental results

In order to evaluate the performance of the proposed system, experiments were performed on the gold standard corpus (see Section 5). We used a standard five-fold cross validation method. As a first step the gold standard data is divided randomly into five parts. Five iterations are then performed. In each of the iterations four different parts of the data (80%) is used to train the classification model, and the remaining part (20%) is used to test the model. The output performances are averaged across the five iterations.

The performances of each of the eight models (see section 4) are measured in terms of precision (*P*), recall (*R*) and F-measure (*F*), which are defined as follows:



where TP refers to *true positive*s or the number of rows to which the classifier correctly assigned the category it recognises (*Title*, *Authors* etc). In the case when that category is incorrectly assigned a *false positive* (FP) is generated, while FN (*false negative*s) represents the number of rows for which the classification model was not able to recognise the category it was constructed for.

Classification experiments were performed in *RapidMiner* (<http://rapid-i.com/content/view/181/190/>), a data mining and machine learning environment. The experiments carried out using standard classification models (see Section 4.2) showed that eight separate *Support Vector Machines* models achieve the best results. Based on that, *feature selection* (using the *backward elimination* algorithm) and parameter optimisation were performed for each of the SVM models.

The performances of the eight SVM models that classify rows as: *Title*, *Authors*, *Affiliation*, *Address*, *Email*, *Abstract*, *Keywords* and *Publication Note* are given in Table 6.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Category*** | ***Precision*** | ***Recall*** | ***F-measure*** |
| *Title* | 99.38% | 98.17% | 98.77% |
| *Authors* | 94.25% | 90.11% | 92.13% |
| *Affiliation* | 90.44% | 89.78% | 90.37% |
| *Address* | 87.50% | 88.11% | 87.80% |
| *Email* | 98.73% | 98.10% | 98.41% |
| *Abstract* | 91.83% | 91.22% | 91.52% |
| *Keywords* | 93.88% | 71.88% | 81.42% |
| *Publication Note* | 90.82% | 83.18% | 86.83% |

Table 6. The performances of the eight SVM classification models.

As can be seen from Table 6, the best performance is achieved for the *Title* category (F-measure of 98.77%) followed by *Email* (F-measure of 98.41%). For the *Authors*, *Abstract* and *Affiliation* categories the performance are somewhat lower with the F-measures of 91.52%, 92.13% and 90.37% respectively. A little more challenging for the classification models were *Address* (F-measure of 87.80%), *Publication Note* (F-measure of 86.83%) and *Keywords* (F-measure of 81.42%).

## Integration of the automatic metadata extraction into CRIS UNS

The proposed system for automatic metadata extraction using *Support Vector Machines* model was integrated in the software system, CRIS UNS. This system supports metadata entry from published scientific research results, as well as storage of publications in digital form, including PDF. The first step in entering metadata, even before uploading the publication in digital form into CRIS UNS, is the selection (or entry) of the proceedings or a journal in which the article is published. After that the curator uploads the publication in PDF format. During the metadata entry process there is an option for "*Automatic extraction of metadata*". Choosing this option triggers automatic metadata extraction from the first page of the PDF by the system presented in this paper. If the data about the proceedings or a journal in which the article is published entered in the first step differ from those extracted by the proposed system, the curator will be warned so that he could perform the appropriate correction. When the curator determines that the metadata is correct he performs the appropriate action to store it into CRIS UNS system database. Since CRIS UNS system data model is CERIF compatible (Ivanović et al., 2011a) the metadata related to the proceedings or a journal in which the article is published is stored using the CERIF data model entities *cfResPubl* and *cfResPubl\_resPubl* (Jörg et al., 2009a). The next step is to prepare the rest of the extracted metadata for entry into CIRS UNS. The data for categories such as *Title*, *Abstract* and *Keywords* are offered to the curator, as extracted by the proposed system. CRIS UNS provides an option of entering the content of an abstract along with its formatting. Given that the extracted metadata contains information about formatting style (see Tables 1-4), a formatted abstract is also offered to the curator. The title of the publication is stored in *cfResPublTitle* entity in the system data model while the abstract and keywords are stored in *cfResPublAbstr* and *cfResPublKey* entities respectively. The next step is processing the results for the: *Authors*, *Address*, *Affiliation* and *Email* categories in order to obtain information on individual authors of the publication being entered. Data obtained in this manner is then used to search the CRIS UNS database of researchers. The database query is performed separately for each of the extracted authors. The CRIS UNS database of researchers is indexed using *Apache Lucene* library (<http://lucene.apache.org/>) which allows *fuzzy search*. This means that the order of the extracted first and last names or any incorrectly recognised characters by the *pdf2html* tool will not change the search results in a significant way. There are three possible outcomes of the query:

* Exactly one researcher matches the query. In this case, the data for that researcher is offered to the curator.
* More than one researcher matches the query. In this case, the curator is presented with the list of researchers found in order to select the appropriate one.
* There aren't any researchers that match the query. In this case the curator is presented with the form for adding new researchers into the database. The form is pre-filled with the extracted metadata used for the query.

The extracted metadata about the authors is entered into *cfPersName* entity. The author's physical and email addresses are stored in *cfPers\_PAddr* and *cfPers\_EAddr* entities. The institution to which the author is affiliated is entered into *cfPers\_OrgUnit* entity in the CERIF data model. All metadata about researchers and publications in CRIS UNS data model are mapped to MARC21 authority and bibliographic records, respectively. The CRIS UNS data model is CERIF compatible in the sense that metadata from MARC21 records can be uniquely mapped to the CERIF entities. This model preserves the existing attributes and references between the CERIF data model entities (Ivanović et al., 2011a).

It is important to note that the automatically extracted metadata in any of the data entry steps is not directly entered into the database, but offered to the curator for possible corrections.

Metadata extraction has been tested on the publications of researchers from the Department of Mathematics and Informatics of the Faculty of Sciences in Novi Sad. The publications are from the fields of Mathematics and Computer Science and come from different conferences and journals and are not part of the corpus presented in Section 5. An analysis of metadata extracted from a sample of 100 papers showed that the performance of the system is in accordance with the results given in Table 6.

## Conclusion

In this paper a system for the fully automated extraction of metadata from scientific publications in PDF format is proposed. The extracted metadata are categorised in eight predefined categories: *Title*, *Authors*, *Affiliation*, *Address*, *Email*, *Abstract*, *Keywords* and *Publication Note*. The extraction task is formalised as a classification problem realised by machine learning methods. Experiments were performed with standard classification models: *Decision Tree*,  *Naive Bayes*, *K-nearest Neighbours* and *Support Vector Machines*. Both a single classifier with all eight categories and eight individual classifiers were tested. Classifiers were evaluated, using five-fold cross validation, on a manually annotated corpus of 100 scientific papers in PDF format, from various conferences, journals and author's personal web pages. Eight separate *Support Vector Machines* models achieved the best performances and based on that, were chosen as a classification model for the proposed system. All of the eight SVM models performed well. The F-measure was over 85% for almost all of the classifiers and over 90% the most of them. Best results were achieved for the *Title* (F-measure of 98.77%) and *Email* (F-measure of 98.41%) categories. Performances for the *Abstract*, *Authors*, *Affiliation* and *Address* categories were slightly lower with F-measures of 91.52%, 92.13%, 90.37% and 87.80%, respectively. Recall values for *Publication Note* and *Keywords* were somewhat lower (between 83% and 71.88%) in comparison to the precision (which is in the range with the rest of the categories) indicating that the current features have not covered all the possible ways by which these metadata are represented (layout, formatting, text content etc) in the scientific papers. Decline in recall for these two categories can also be explained by their low frequency in the gold standard corpus (2,38% for *Publication Note* and 0,98% for *Keywords*) i.e. training sets for the models.

The proposed system for automatic metadata extraction using *Support Vector Machines* model was integrated in the software system, CRIS UNS. Metadata extraction has been tested on the publications of researchers from the Department of Mathematics and Informatics of the Faculty of Sciences in Novi Sad. An analysis of the extracted metadata showed that the performances the system for the previously unseen data are in accordance with tose obtained by the cross validation from eight separate SVM classifiers.

Further studies are planned in order to improve performance e.g. by using hand-crafted rules and new features to boost recall for the *Keywords* and *Publication Note* categories. Since the first step in entering metadata into CRIS UNS is the selection of the proceedings or a journal in which the article is published, this information will be used to construct classification models with better performances. Further processing of the extracted results for some of the categories, using methods from the field of IE, will be performed. For example *Publication Note* metadata will be analysed in order to extract detailed information such as date, publisher, conference venue, and the volumes of journals and so on. Experiments will also be performed with the methods of semi-supervised learning such as co-training with a goal of overcoming the problem of manual annotation of a large number of scientific papers.

## Acknowledgment This paper is part of the research project “Infrastructure for Technology Enhanced Learning in Serbia” supported by the Ministry of Education and Science of the Republic of Serbia [Project No. 47003].

## References

Abugessaisa, I., (2010), "Geospatial metadata extraction from product description document applying methods from ontology engineering", *International Journal of Metadata, Semantics and Ontologies*, Vol. 5 No 4., pp. 321-332.

Adams, J., (2009), "The use of bibliometrics to measure research quality in UK higher education institutions", *Archivum Immunologiae et Therapiae Experimentalis*, Vol. 57 No. 1, pp. 19-32.

Anderson, J.D. and Pérez-Carballo, J., (2001), "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort", *Information Processing & Management*, Vol. 37 No. 2, pp. 255-277.

Asserson, A., Jeffery, K. and Lopatenko, A., (2002). "CERIF: Past, Present and Future: An Overview", in *Proceedings of the 6th International Conference on Current Research Information Systems, University of Kassel, August 29 - 31, 2002,* pp. 30-40.

Belić, K. and Surla, D., (2008a), "User-friendly web application for bibliographic material processing", *The Electronic Library*, Vol. 26 No. 3, pp. 400-410.

Belić, K. and Surla, D., (2008b), "Model of a user friendly system for library cataloguing", *Computer Science and Information Systems*, Vol. 5 No. 1, pp. 61-85.

Bergmark, D., (2000), "Automatic extraction of reference linking information from online documents", *Cornell University, Ithaca, NY*.

Boberić, D. and Surla, D., (2009), "XML editor for search and retrieval of bibliographic records in the Z39.50 standard", *The Electronic Library*, Vol. 27 No. 3, pp. 474-495.

Cardie, C., (1993), "A case-based approach to knowledge acquisition for domain-specific sentence analysis", in *Proceedings of the eleventh national conference on Artificial intelligence*. *AAAI'93*, *July 11-15,* *1993,* AAAI Press, pp. 798-803.

Chieu, H.L. and Ng, H.T., (2002), "A maximum entropy approach to information extraction from semi-structured and free text", in *Proceedings of the Eighteenth national conference on Artificial intelligence*, *Menlo Park*, *CA*, *USA,* *July 28 - August 1, 2002,* American Association for Artificial Intelligence, pp. 786-791.

Cristianini, N., (2000), *An introduction to support vector machines: and other kernel-based learning methods,* Cambridge, New York, Cambridge University Press.

Crystal A. and Land, P., (2003), "Metadata and Search: Global Corporate Circle DCMI 2003 Workshop", available at: http://www.dublincore.org/groups/corporate/Seattle/ (accessed 14 January 2011).

Cui, B., (2009), "Scientific literature metadata extraction based on HMM", in *Proceedings of the 6th international conference on Cooperative design, visualization, and engineering*. *CDVE'09, Luxembourg, Luxembourg, September 20 - 23, 2009,* Springer-Verlag, pp. 64-68.

Cui, B. and Chen, X., (2010), "An Improved Hidden Markov Model for Literature Metadata Extraction", in *Proceedings of the 6th international conference on Advanced intelligent computing theories and applications: intelligent computing. ICIC'10, Changsha, China, August 18-21, 2010,* Springer-Verlag, pp. 205-212.

Devignes, M. et al., (2010), "BioRegistry: Automatic extraction of metadata for biological database retrieval and discovery", *International Journal of Metadata, Semantics and Ontologies,* Vol 5. No. 3, pp.184–193.

Dimić, B., Milosavljević, B. and Surla, D., (2010), "XML schema for UNIMARC and MARC 21", *The Electronic Library*, Vol. 28 No. 2, pp. 245-262.

Dimić, B. and Surla, D., (2009), "XML editor for UNIMARC and MARC 21 cataloguing", *The Electronic Library*, Vol. 27 No.3, pp. 509-528.

Domingos, P. and Pazzani, M., (1997), "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss", *Mach. Learn.*, Vol. 29, pp. 103-130.

Finkel, J.R., Grenager, T. and Manning, C., (2005), "Incorporating non-local information into information extraction systems by gibbs sampling", in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* *University of Michigan, USA, June 25 - 30, 2005,*  pp. 363-370.

Flynn, P. et al., (2007), "Automated template-based metadata extraction architecture", in *Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers*. *ICADL'07, Hanoi, Vietnam, December 10 - 13, 2007*, Springer-Verlag, pp. 327-336.

Giuffrida, G., Shek, E.C. and Yang, J., (2000), "Knowledge-based metadata extraction from PostScript files", in *Proceedings of the fifth ACM conference on Digital libraries*. *DL '00*, *San Antonio, TX, USA, June 02 - 07, 2000,* ACM, pp. 77-84.

Greenberg, J., Spurgin, K. and Crystal, A., (2006), "Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions", *International Journal of Metadata, Semantics and Ontologies,* Vol. No 1., pp.3–20.

Groza, T., Handschuh, S. and Hulpus, I., (2009), "A document engineering approach to automatic extraction of shallow metadata from scientific publications", *Galway: DERI Galway, National University of Ireland*, available at: http://cogprints.org/5859/1/Thesis-David-Nadeau.pdf (accessed 4 February 2011)

Han, H. et al., (2003), "Automatic document metadata extraction using support vector machines", in *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*. *JCDL '03, Houston, TX, USA, May 27 - 31, 2003*, IEEE Computer Society, pp. 37-48.

Han, H. et al., (2005), "Rule-based word clustering for document metadata extraction", in *Proceedings of the 2005 ACM symposium on Applied computing*. *SAC '05*, *Santa Fe, New Mexico, March 13 - 17, 2005,* ACM, pp. 1049-1053.

Heidorn, P.B. and Wei, Q., (2008), "Automatic metadata extraction from museum specimen labels", in *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications, Berlin, September 22 - 26, 2008*, Dublin Core Metadata Initiative, pp. 57-68.

Hong-ye, C., (2009), "Method of Web Information Extraction Based on Decision Tree", in *Proceedings of the 2009 International Forum on Information Technology and Applications - Volume 01*, *China, May 15-17, 2009,* IEEE Computer Society, pp. 664-666.

Hu, Y. et al., (2006), "Automatic extraction of titles from general documents using machine learning", *Information Processing & Management*, Vol. 42 No. 5, pp. 1276-1293.

Ivanović, D., Milosavljević, G., Milosavljević, B. and Surla, D., (2010), "A CERIF-compatible research management system based on the MARC21 format", *Program: Electronic libarary and information systems*, Vol. 44, No. 3, pp. 229-251

Ivanović, D., Surla, D. and Racković, M., (2011a), "A CERIF data model extension for evaluation and quantitative expression of scientific research results", *Scientometrics*, Vol. 86 No. 1, pp. 155-172.

Ivanović, D., Surla, D. and Konjović, Z, (2011b), "CERIF compatible data model based on MARC21 format", *The Electronic Library*, Vol. 29, No. 1, pp. 52-70.

Jörg, B., Krast, O., Jeffery, K. and Grootel, G., (2009a), "CERIF 2008 - 1.0 Full Data Model (FDM) Introduction and Specification", available at: http://www.eurocris.org/fileadmin/cerif-2008/CERIF2008\_1.0\_FDM.pdf (accessed 4 September, 2010)

Jörg, B., Krast, O., Jeffery, K. and Grootel, G., (2009b), "CERIF 2008 - 1.0 XML Data Exchange Format Specification", available at: http://www.eurocris.org/fileadmin/cerif-2008/CERIF2008\_1.0\_XML.pdf (accessed 4 September, 2010)

Klink, S., Dengel, A. and Kieninger, T., (2000), "Document structure analysis based on layout and textual features", in *Proceedings of the International Workshop on Document Analysis Systems, DAS2000*, *Boston, MA, USA, June 9 -11, 2010,* pp. 99-111.

Larkey, L.S. and Croft, W.B., (1996), "Combining classifiers in text categorization", in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '96, Zurich, Switzerland, August 18 - 22, 1996* ACM, pp. 289-297.

Lattner, A.D. and Herzog, O., (2003), "Instance-based learning and information extraction for the generation of metadata", in *Proceedings of the 3rd International Conference on Knowledge Management. I-KNOW 03,* *Gaithersburg, MD, USA November 29 - December 02, 1994*, pp. 472-479.

Lawrence, S., Giles, L. and Bollacker, K., (2002), "Digital libraries and autonomous citation indexing", *Computer*, Vol. 32 No. 6, pp. 67-71.

Liddy, E.D. et al., (2002), "Automatic metadata generation & evaluation", in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, *Tampere, Finland, August 11 - 15, 2002,* pp. 401-402.

Lin, S. et al., (2010), "Extracting formulaic and free text clinical research articles metadata using conditional random fields", in *Proceedings of the The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. *NAACL HLT 2010, Second Louhi Workshop on Text and Data Mining of Health Documents*, *Los Angeles, USA, June 1-6, 2010*, pp. 90–95.

Mao, S., Kim, J.W. and Thoma, G.R., (2004), "A Dynamic Feature Generation System for Automated Metadata Extraction in Preservation of Digital Materials", in *Proceedings of the First International Workshop on Document Image Analysis for Libraries. DIAL'04*, *Palo Alto, California, USA, January 23 - 24*, IEEE Computer Society, pp. 225-232.

Marinai, S., (2009), "Metadata Extraction from PDF Papers for Digital Library Ingest", in *Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain*,*July 26 - 29, 2009,* pp. 251-255.

McCallum et al., (2000), "Automating the Construction of Internet Portals with Machine Learning", *Inf. Retr.*, Vol. 3, pp. 127-163.

McCallum, A., Freitag, D. and Pereira, F., (2000), "Maximum entropy Markov models for information extraction and segmentation", in *Proceedings of the Seventeenth International Conference on Machine Learning*, *Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pp. 591-598.

Milosavljević, B., Boberić, D. and Surla, D., (2010), "Retrieval of bibliographic records using Apache Lucene", *The Electronic Library*, Vol. 28 No. 4, pp. 525-539.

Milosavljević B. and Tešendić, D., (2010), "Software architecture of distributed client/server library circulation system", *The Electronic Library*, Vol. 28 No. 2, pp. 286-299.

Milosavljević, G., Ivanović, D., Surla, D. and Milosavljević, B. (2011), "Automated Construction of the User Interface for a CERIF-Compliant Research Management System", *The Electronic Library,* Vol. 29, No. 5 (in press)

Nadeau, D. and Sekine, S., (2007), "A survey of named entity recognition and classification", *Lingvisticae Investigationes*, Vol. 30 No. 1, pp. 3-26.

Peng, F. and Mccallum, A., (2004), "Accurate Information Extraction from Research Papers Using Conditional Random Fields", in *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics. HLTNAACL, Boston, MA, USA, May 2-7, 2004,* pp. 329-336.

Quinlan, J.R., (1986), "Induction of Decision Trees", *Mach. Learn.*, Vol. 1, pp. 81-106.

Quinlan, J.R., (1987), "Simplifying decision trees", *Int. J. Man-Mach. Stud.*, Vol. 27, pp. 221-234.

Rađenović, J., Milosavljević, B. and Surla, D., (2009), "Modelling and implementation of catalogue cards using FreeMarker", *Program: Electronic Library and information systems*, Vol. 43 No. 1, pp. 62-76.

Ramshaw, L.A. and Marcus, M.P., (1995), "Text chunking using transformation-based learning", in *Proceedings of the Third ACL Workshop on Very Large Corpora,* *Boston, MA, June 30, 1995,* pp. 82-94.

Schwartz, C., (2001). *Sorting out the Web: approaches to subject access*, Westport Conn., Ablex Pub.

Sekine, S. and Grishman, R., (1998), "A Decision Tree Method for Finding and Classifying Names In Japanese Texts", in *Proceedings of the Sixth Workshop on Very Large Corpora*, *Montreal, Quebec, Canada, August 15 - 16, 1998*.

Takasu, A., (2003), "Bibliographic attribute extraction from erroneous references based on a statistical model", in *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*. *JCDL '03*, *Houston, TX, USA, May 27 - 31, 2003*, IEEE Computer Society, pp. 49-60.

Takeuchi, K. and Collier, N., (2002), "Use of support vector machines in extended named entity recognition", in *Proceeding of the 6th conference on Natural language learning. COLING-02*, *Taipei, August 27, 2002,* pp. 1-7.

Tešendić, D., Milosavljević, B. and Surla, D., (2009), "A library circulation system for city and special libraries", *The Electronic Library*, Vol. 27 No. 1, pp. 162-186.

Vapnik, V., (1998), *Statistical learning theory*, Wiley, New York, NY.

Vidaković, M. Milosavljević, B., Konjović, Z. and Sladić, G., (2009), "Extensible java EE-based agent framework and its application on distributed library catalogues", *Computer Science and Information Systems*, Vol. 6 No. 2, pp. 1-28.

Yang, Y., (2001), "A study of thresholding strategies for text categorization", in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, *SIGIR '01,New Orleans, LA, USA, September 9 - 13, 2001* ACM, pp. 137-145.

Yilmazel, O., Finneran, C.M. and Liddy, E.D., (2004), "Metaextract: an NLP system to automatically assign metadata", in *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. *JCDL '04, Tucson, AZ, USA, June 07 - 11, 2004,* ACM, pp. 241-242.

Yin, P. et al., (2005), "Metadata Extraction from Bibliographies Using Bigram HMM", in *Z. Chen et al., eds.* *Digital Libraries: International Collaboration and Cross-Fertilization*. *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 1-14.

Zimmerman, E., (2002), "Cris-cross: Current research information systems at a crossroads", in *Adamczak, W., Nase, A. (Eds.) (2002), Gaining Insight,from Research Information. Proceedings of the 6th International Conference on Current Research Information Systems, University of Kassel, August 29 - 31,* pp.29-3l.