Rule-based extraction of metadata from scientific articles in Serbian language: a case study in YUINFO conference proceedings

Petrović Gajo, Kovačević Aleksandar, Konjović Zora

University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

Abstract: *Since English is a predominant language in scientific publishing a great majority of information retrieval systems is oriented towards it. An effort is therefore needed to index articles written in other languages, especially ones spoken in relatively small countries such as Serbia. As an important step in the indexing process we present a rule-based approach to the extraction of metadata from scientific articles in Serbian language. A corpus of articles from the well established Serbian computer science conference YU INFO was collected. We than developed a set of rules using different types of features (positioning, orthographic, presence of keywords etc.). We extracted three metadata categories: Title, Author (containing author's full names) and Abstract. The extraction component was evaluated for Title and Author metadata categories on gold standard set of metadata automatically collected from the conference web site. The F-measure for the Title category was 76% (with the precision of 68%), and 81% (with the precision 72%) for the Author category. We also present the results of our initial insight of the contents of the article abstracts.*

1. INTRODUCTION

In recent years we are witnessing an explosion of digital publishing, a scientific paper is published every 3 minutes [1]. Consequently a myriad of systems (Google Scholar [2], CiteSeerX [3], ACM [4], IEEE [5], DSpace [6] and CRIS UNS [7]) and initiatives (Semantic Web [8]) emerged with the goal of indexing, curating and providing intuitive search to scientific articles. Majority of these systems only index papers in English language, thus leaving scientific forums (especially of older date) published in other languages omitted from the curation process. An effort is therefore needed in order to document and organize valuable knowledge published in other languages (mostly in national venues).

In this paper we address an important step of the curation process by proposing a system for the extraction of metadata from scientific papers published in a well established Serbian computer science conference – YU INFO [9]. Our system is rule-based and the following metadata categories are extracted:

- Serbian and English titles
- List of authors' full names
- Serbian and English abstract text

We evaluated our system on the corpus of 1256 papers published in previous YU INFO conferences held from 2006 to 2012, and present some initial insight about the obtained metadata.

The paper is organized as follows: section 2 presents related work in the field, section 3 explains the dataset used in this paper, section 4 describes the methodology, section 5 shows the parsing results and section 6 presents a possible use of the extracted data.

2. RELATED WORK

The problem of metadata extraction was addressed before in both rule-based approaches [10] as well as machine learning based ones [11,12]. While machine learning based ones can be more adaptable and robust than rule based approaches, they require large manually annotated training sets. Rule based approaches can offer good performance at the expense of time required to construct them [13]. They usually make use of the format scientific papers are usually written in, like was done in [14] where rules such as "title is usually found in the first parts of the text and has the largest font" are used to extract metadata elements from PostScript files. These rules are often written in the form of regular expressions as was done in [15], and they can be made to obtain common metadata such as: title, author, sections and references. Once metadata is obtained, it can be used in a variety of analysis which can lead to interesting results as shown in [16], where co-authorship relation was analyzed.

Since we focused on papers from a particular conference (YU INFO) with predefined formatting guidelines, adaptability and robustness weren't our primary goals. Thus we opted for a rule based approach.

3. THE YU INFO PROCEEDINGS CORPUS

The YU INFO conference was founded in 1995 and has been held annually since. The papers are submitted in the PDF format and are written based on the document format[1] mandated by the conference organizers. This format gives specific instructions on how to write important sections of the paper, including:

- Title: font, case and positioning
- Author and co-author: font, position, name separation, affiliations and e-mail addresses.
- Abstract: font, positioning and title.
- Section: font, separation, numbering.
- References: font and format.

The main intent of these formatting rules is to increase article readability, rather than to achieve automated document parsing capabilities. In fact, it is not unusual to require authors to separately, manually input some of the previously mentioned fields when submitting the paper, which tends to be a time-consuming and error-prone process.

Our corpus comprised papers in PDF format collected from the publicly available repository of accepted papers in the span from 2006 to 2012 (http://www.e-drustvo.org/yuinfo/zbornici.html). The gold standard paper metadata, used to measure the performance of the extraction component, was obtained from the official Web based YU INFO repositories (available in HTML format). Papers are organized by year and for each year; they are

[1] http://www.informacionodrustvo.org/dotAsset/10771.pdf.

categorized by the scientific discipline[2]. Each entry contains the name of the academic discipline and a link to a page with the list of papers published that year for that particular discipline. The list of papers includes the list of authors, the paper title in Serbian (or English if the entire paper was written in English) and a link to the paper in PDF format. The corpus statistics are given in Table 1.

Total papers		1256
Total authors and co-authors		2933
Unique authors and co-authors		2143
Papers by scientific discipline	Networking	225
	Artificial Intelligence and computer simulation	105
	Software and tool development	157
	Hardware	42
	Software use in military and security applications	48
	Applied informatics	266
	Data protection and security	43
	Information systems	153
	e-Society and the Internet	217

Table 1. Corpus statistics

A Web crawler, used to collect our corpus was implemented in the Python programming language using the *mechanize* [17] library. A script program written in Bash utilizing *sed* and *grep* UNIX tools was used to extract Title and Author metadata categories and URLs for the PDF files from the obtained HTML pages.

4. METHODOLOGY

The goal of automatic metadata extraction is to obtain key metadata from paper's PDF file format representation as they were submitted to the conference or journals.

Our proposed extraction method is shown on Figure 1.

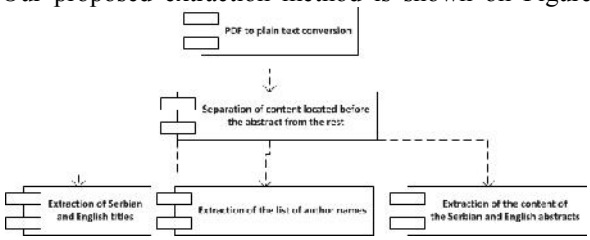

Figure 1. Metadata extraction process

The "PDF to plain text conversion" component uses the *pdftotext* program, which is a part of the *Poppler library* [18]. The program takes a PDF file as input and produces a plain text file containing the text of the PDF file as output. This tool was applied for each file.

The second step performs separation of the text into two parts: one containing the titles and a list of authors' names, and the other containing the remaining text of the paper.

The "Extraction of Serbian and English titles" process extracts the two titles (Serbian and English) from the first part of the text separated in the previous step. It is assumed that these fields have been written at the top of the document.

The "Extraction of the list of author names" component extracts a list of authors' names from the first part of the separated text. This is done by dividing the comma separated names from the first line to meet a certain criteria, given as a regular expression in the form: `.*(.[a-zčćšđž"]{2,}\\s*,)+.*` .

The content of article abstracts was obtained in the "Extraction of the content of the Serbian and English abstracts" step. Abstracts were detected as a passage of text preceded by particular keywords (e.g. ***Abstract***, ***Apstrakt***, and their variations) and succeeded by the beginning of the first section of the paper (or the abstract written in the other language).

We designed a set of pattern matching rules based on: position, case, alphabet (Serbian Cyrillic, Serbian Latin or English Latin) and the presence of certain keywords or regular expressions. The total amount of rules used for each metadata type is given in table 2.

	Rules used
Author names	4
Abstracts	6
Titles	2

Table 2. Rule amount per metadata type.

Most of these rules were regular expressions (implemented in Java) which matched text content of a line, with just one rule for title extraction using line position in the document.

The output of our extraction pipeline was the set of the following metadata fields: Serbian and English titles, list of authors' names and Serbian and English abstracts.

5. RESULTS

We evaluated our extraction pipeline by comparing the titles and author names extracted from PDF files with the gold standard values obtained from the HTML pages.

The evaluation against the gold standard was done for all files that produced a valid plain text file (1244 in total). The performance was measured in terms of precision (P), recall (R) and F-measure (F), which are defined as follows:

$$Recall = \frac{tp}{tp + fp} \qquad Precision = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

where *tp*, *fp* and *fn* denote true positive, false positive and false negative cases.

The extracted metadata fields were compared to the gold standard values using the Levenshtein string matching distance [19] in order to compensate for the small differences in values caused by the errors of the *pdf2text* tool (incorrect parsing of Serbian Latin characters: š, č, ć, ž, đ). The values of two or less for the Levenshtein distance were considered as match. In cases where a paper contained titles written in both Serbian and English, the

[2] The list of disciplines for a given year can be found at *http://www.e-drustvo.org/proceedings/YuInfo$YEAR/html/proceedings.html* ($YEAR is the parameter which replaces the desired year, e.g. 2006).

Serbian title was used in compared to the gold standard, while the English title would only be used if no Serbian title was found (i.e., papers written entirely in English). The results for the extraction of *Author* and *Title* metadata fields are given in Table 3.

	Precision	Recall	F-measure
Title	68%	85%	76%
Author	72%	91%	81%

Table 3. Results for the extraction of *Author* and *Title* metadata fields

The overall results show that the *Author* category (81% F-measure) outperformed the *Title* category (76% F-measure). Precision values were relatively low for both *Title* (68%) and *Author* (72%) categories, while both had high recall (85% for *Title* and 91% for *Author*).

In order to explain the performance of our rule-based extraction component we analysed a random sample of 100 papers and identified two types of errors: fragmented text caused by *pdf2text* conversion tool (approximately 30% of errors fall into this group), and papers written outside of the conference formatting guidelines (about 15% of errors are of this type).

5. ANALYSIS OF THE EXTRACTED METADATA

In order to gain an initial insight about the knowledge published in the conference we extracted three most frequent unigrams and bigrams from both Serbian and English abstracts. Frequencies were counted per article i.e., all the mentions of the same term in an article were grouped and counted as one. For example if the term "neural network" was mentioned 30 times in an article it's per article frequency would still be 1. Terms were extracted and organized twofold: by year and by scientific discipline. The term generation process is shown on Figure 2.

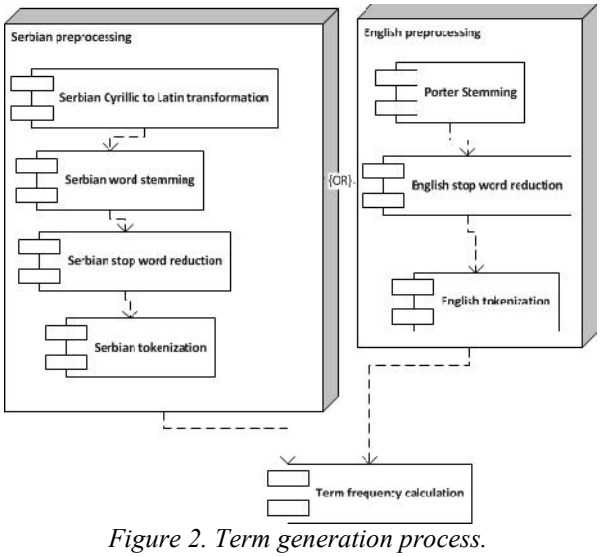

Figure 2. Term generation process.

Abstracts were preprocessed separately depending on the language. The first step in both cases was tokenization where the texts were splits into separate tokens. In the case of Serbian text, this step also included conversion from Serbian Cyrillic to Serbian Latin alphabet. The conversion was necessary because further processing steps were created to be able to only handle input in Latin alphabet. Stemming (reducing words to their base form) was preformed as the next step in order to reduce the lexical variability i.e. improve the quality of the extracted term frequency lists. We used the well known Porter stemmer for English texts, while for Serbian texts were stemmed with our modified version of the Croatian language stemmer[3]. As a final step stop (non content-bearing) words were removed. We used a hand crafted list of Serbian and English stop words. The preprocessing for both languages was implemented using the *Apache Lucene* library [20].

Tables 4 and 5 present the top three extracted unigrams and bigrams organized by year and scientific discipline respectively.

By observing the unigrams given in Table 4 we can see that they generally represent common conference topics (networks, services, etc.), while bigrams suggest yearly topic tendencies: 2006-2008 seemed to focus on IT topics, 2009-2010 had an increase in papers about open source technologies and 2011-2012 had papers published in the field of web technologies and wireless communication.

Unigrams in Table 5 are mostly keywords that are good representatives of the respective scientific disciplines, while bigrams are good indicators of the most popular sub-disciplines i.e., the ones with the high amount of published articles.

6. CONCLUSION

In this paper we have proposed a system for the extraction of metadata from scientific publications in PDF format written in Serbian language. In particular, we directed our efforts towards articles published between 2006 and 2012 in the well established Serbian computer science conference called YUINFO. To the best of our knowledge, this work is one of the first attempts that concentrates on metadata in Serbian language.

Our approach was rule-based and extracted the following metadata categories: *Title*, *Author* and *Abstract*. Rules were developed in order to exploit the relative structure of metadata mandated by the conference guidelines. The system was evaluated on a gold standard corpus comprising title and author metadata available in HTML format on the conference Web site. Results were promising, with the F-measure of 81% for *Author* category and 76% for *Title* category. We have also identified two major sources of errors: low quality output of the PDF to plain text conversion tool and texts formatted outside the mandated guidelines. As a first glance into our corpus we extracted top three unigrams and bigrams and organised them by year and scientific discipline.

Future work includes improving the PDF to plain text conversion step of our system; developing more robust rule to encompass metadata from poorly formatted papers and applying various data mining methods on the extracted metadata such as clustering, association rules etc.

[3] http://www.nljubesic.net/resources/tools/stemmer-for-croatian/

		Top unigram terms by occurrence per paper							Top bigram terms by occurrence per paper						
Year	Lang	Term	#	Term	#	Term	#	Total words	Term	#	Term	#	Term	#	Total words
2006	EN	service	37	generic	27	network	27	17891	high level	6	information technology	6	computer simulation	5	19077
2006	RS	promovisati	28	tehnologija	24	korisnik	23	7905	procentualna razmera	9	mobilni telefon	6	sistem upravljanja	4	8882
2007	EN	busy	30	tehnology	23	network	23	7868	information technology	6	developing countries	4	decision making	4	9489
2007	RS	promovisati	24	elektronsko	23	korišćenje	23	7537	sistem kontrole	5	sistem podrške	5	informacione tehnologije	4	9106
2008	EN	service	26	program	26	network	21	7975	information technology	5	program package	9	management system	5	9695
2008	RS	korišćenje	26	promovisati	24	web	23	7489	programski paket	9	programski alfabet	6	informacione tehnologije	6	9160
2009	EN	manage	27	represent	25	program	24	7832	open source	6	computer network	4	information technology	4	9655
2009	RS	promovisati	28	upravljanje	25	prikaz	23	7164	sistem upravljanja	6	programski alfabet	5	vojna akademija	4	8353
2010	EN	manage	27	service	25	network	24	7019	open source	9	software package	8	web service	6	8766
2010	RS	promovisati	29	primena	25	skor	22	7226	programski paket	9	programski alfabet	5	elektronski sistem	4	8495
2011	EN	service	27	network	26	web	22	5247	web service	7	network sensor	7	wireless sensor	5	6850
2011	RS	korišćenje	27	zahtev	24	prikaz	22	7965	web servis	6	web aplikacija	6	bežični senzor	5	10206
2012	EN	busy	22	technology	22	network	22	7668	large number	6	wireless sensor	5	data mining	4	9102
2012	RS	tehnologija	24	skor	22	poslovno	20	6189	sistem upravljanja	5	bežični senzorski	4	nove tehnologije	3	7208

Table 4. Top 3 unigrams and bigrams by occurrence per paper grouped by year and language, ("#" denotes the count of term occurrences)

		Top unigram terms by occurrence per paper							Top bigram terms by occurrence per paper						
Scientific discipline	Lang	Term	#	Term	#	Term	#	Total words	Term	#	Term	#	Term	#	Total words
Networks	EN	network	93	service	51	mobile	35	11836	network sensor	9	wireless sensor	9	simulation result	6	13372
Networks	RS	servis	36	mobilni	35	karakteristika	31	7866	mobilni telefon	9	bežični senzorski	9	verovatnoća greške	6	8973
AI & Simulation	EN	simulation	26	algorithm	21	optimal	13	3743	neural network	6	computer simulation	6	expert system	5	4905
AI & Simulation	RS	skor	22	primena	18	simulacija	17	5048	simulacioni model	5	mobilni robot	4	sistem upravljanja	4	6498
Software	EN	web	33	program	21	user	20	6668	web application	9	open source	7	web service	5	8383
Software	RS	promovisati	30	web	29	upravljanje	25	7030	web aplikacija	11	programski jezik	10	sistem upravljanja	7	8801
Hardware	EN	measure	11	sensor	8	test	8	1801	under linux	3	benchmark software	3	network sensor	3	2214
Hardware	RS	korišćenje	9	realizacija	7	skor	6	2089	linux operativni	4	kontrolni sistem	3	poređenje performansi	3	2722
Security	EN	military	14	operation	8	communication	7	2065	military academy	5	tactical units	2	information system	2	2446
Security	RS	komunikacija	10	vojno	10	promovisanje	9	2013	vojna akademija	4	ministarstvo odbrane	3	sistem odbrane	2	2299
Applied Software	EN	program	41	import	39	work	37	11664	software package	10	finite element	10	program package	8	14241
Applied Software	RS	promovisati	35	korišćenje	33	paket	31	10727	programski paket	21	mašinski fakultet	6	infomacione tehnologije	5	12378
Data protection	EN	security	16	web	7	service	7	1719	public key	4	access control	3	information technology	3	2132
Data protection	RS	bezbednost	9	kontrola	6	napad	6	1467	informacione tehnologije	3	kontrola pristupa	3	bezbednost pravosuđa	2	1717
Information system	EN	busy	31	integration	31	service	26	7894	geographical information	9	management system	9	information system	9	9177
Information system	RS	potreba	29	promovisati	29	poslovno	29	6164	poslovni proces	12	sistem upravljanja	8	geografska informacija	8	7187
Internet E-Society	EN	learn	46	technology	43	web	41	14110	information communication	14	information technology	11	communication technology	9	15811
Internet E-Society	RS	elektronski	52	učenje	35	tehnologija	34	9071	elektronsko učenje	9	informacione tehnologije	8	sistem učenja	7	10834

Table 5. Top 3 unigrams and bigrams by occurrence per paper, grouped by scientific discipline and language, ("#" denotes the count of term occurrences)

REFERENCES

[1] J. DeShazo, D. LaVallie and M. Wolf, "*Publication trends in the medical informatics literature: 20 years of medical informatics*", BMC Med, 2009.

[2] "Google Scholar," [Online]. Available: http://scholar.google.com/. [Accessed 22 1 2013].

[3] "CiteSeerX," [Online]. Available: http://citeseerx.ist.psu.edu/. [Accessed 22 1 2013].

[4] "ACM," [Online]. Available: http://www.acm.org/publications. [Accessed 22 1 2013].

[5] "IEEE," [Online]. Available: https://www.ieee.org/publications_standards. [Accessed 22 1 2013].

[6] "DSpace," [Online]. Available: http://dspace.mit.edu/. [Accessed 22 1 2013].

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

[8] "Semantic Web W3C," [Online]. Available: http://www.w3.org/standards/semanticweb/. [Accessed 22 1 2013].

[9] "YuInfo," [Online]. Available: http://www.e-drustvo.org/yuinfo/. [Accessed 22 1 2013].

[10] E. Liddy, E. Allen, S. Harwell, S. Corieri, O. Yilmazel, N. Ozgencil, A. Diekema, N. McCracken and J. Silverstein, "*Automatic metadata generation & evaluation*", 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, 2002.

[11] A. Kovačević, D. Ivanović, B. Milosavljević, Z. Konjović and D. Surla, "*Automatic extraction of metadata from scientific publications for CRIS systems*", Program: Electronic Library and Information Systems, vol. 45, no. 4, pp. 376-396, 2011.

[12] A. Kovačević, D. Ivanović, B. Milosavljević, Z. Konjović and G. Nenadić, "*Mining methodologies from NLP publications: A case study in automatic terminology recognition*", Computer Speech & Language, vol. 26, pp. 105-126, 2012.

[13] S. Klink, A. Dengel and T. Kieninger, "*Document structure analysis based on layout and textual features*", International Workshop on Document Analysis Systems, Boston, 2011.

[14] G. Giuffrida, E. Shek and J. Yang, "*Functionalities for automatic metadata*" International Journal of Metadata, Semantics and Ontologies, vol. 1, no. 1, pp. 3-20, 2006.

[15] A. Ojokoh, S. Adewale and O. Falaki, "*Automated document metadata extraction*", Journal of Information Science, vol. 35, no. 5, pp. 563-70, 2009.

[16] M. Radovanović, J. Ferlež, D. Mladenić, M. Grobelnik and M. Ivanović, "*Mining and Visualizing Scientific, Publication Data from Vojvodina*", Novi Sad Journal of Mathematics, vol. 37, no. 2, pp. 161-180, 2007.

[17] "Mechanize, a Python library for web crawling," [Online]. Available: http://wwwsearch.sourceforge.net/mechanize/. [Accessed 22 1 2013]

[18] "Poppler, an open source PDF rendering library," [Online]. Available: http://poppler.freedesktop.org/. [Accessed 22 1 2013].

[19] V. I. Levenshtein, "*Binary codes capable of correcting deletions, insertions, and reversals*", Soviet Physics Doklady, vol. 10, p. 707, 1966.

[20] "Apache Lucene," [Online]. Available: http://lucene.apache.org/core/. [Accessed 22 1 2013].

ACKNOWLEDGEMENT

Research presented in this paper is partly funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Grant No. III 47003.