A Strategy for Automatically Extracting References from PDF Documents

Neide Ferreira Alves
Universidade do Estado do Amazonas
Manaus, Brazil
nfalves@uea.edu.br

Rafael Dueire Lins
Universidade Federal de Pernambuco
Recife, Brazil
rdl@ufpe.br

Maria Lencastre
Universidade de Pernambuco
Recife, Brazil
mlpm@ecomp.poli.br

***Abstract* — Every day the number of citations an author receives is becoming more important than the size of his list of publications. The automatic extraction of bibliographic references in scientific articles is still a difficult problem in Document Engineering, even if the document is originally in digital form. This paper presents a strategy for extracting references of scientific documents in PDF format. The scheme proposed was validated in LiveMemory platform, developed to generate digital libraries of proceedings of technical events.**

Keywords - information extraction, bibliographic references, document processing, regular expression, learning.

I. INTRODUCTION

The acknowledgement of the sources of a technical article is in its list of bibliographical references. Conversely, the number of citations a given article receives may be an indication of its importance in a given area. Thus, citation indices are becoming more important than the size of the list of publications of a given author or researcher. Collecting such information is far from being a trivial task, however.

In the case of legated paper documents an effort of paramount dimension is necessary. This is due to the need to either re-type such data or to scan the document, to automatically process it, in order to enhance the quality of the image, attempt to find the list of references, and finally transcribe it via OCR [6]. Such scheme is still processing intensive and error prone. In the case of electronically generated documents of formats such as PDF, PS, HTML and XML, the task of reference spotting is much easier, and tends to be more accurate than in the case of legated printed ones. This does not mean that it is a straightforward task. The automatic extraction of references is still a difficult problem in Document Engineering.

In proceedings, neither authors use, nor editors check to guarantee that the adopted bibliographic templates were strictly followed. Problems often arise in the items in the list of references, such as: incompleteness, existence of different formats out of the pattern, abbreviations, etc.

This work details a process for extracting bibliographical references in the context of the LiveMemory Project [9]. LiveMemory is a platform developed for the semi-automatic generation of digital libraries of proceedings of technical events. It allows processing the image of scanned documents (JPEG, TIFF e PNG), automatic indexation of files, extraction and storage of information in databases, such as: paper title, authors and their institutions, keywords, abstracts and references, year of publication, etc. The platform also allows the generation of reports about most used keywords, most cited references, etc.

This paper presents the strategy used in the LiveMemory platform for extracting the list of references of a PDF document, which was digitally generated. Regular expressions, together with classification and identification based on K-NN algorithm and the Naïve Bayes algorithm were used for this purpose. The whole process is presented throughout this paper, which is organized in the following way: Section 2 presents related work in the literature; Section 3 details the strategy for extracting references; Section 4, presents the extraction system, which gives support for the proposal; Section 5 presents some performance tests; Finally, Section 6 presents the conclusions and draws lines for future work.

II. RELATED WORK

Several researchers proposed ways of extracting information from bibliographical references. This section describes some of such work and also tools that are close to our proposal.

The first work on bibliographic reference extraction used the Hidden Markov Model (HMM) technique [12]. In reference [5] the authors consider the tagging process for classifying the items that compose references, and also the automatic induction of a set of rules for extracting specific features. Reference [2] extracts information from texts in Japanese using OCR. First, blocks are labeled with title, abstract and references; after each block is re-labeled for the extraction of information that one requires.

In reference [7] the authors propose the extraction of the names of authors from academic papers, using the identification of uppercase letters, lines breaks, tagging of characters and use of regular expressions. Aljaber and his colleagues [3] use the scope of the citations to verify the similarity between texts and the partition into classes for applying the K-Means algorithm. In [4] a combination of regular expressions, a system based on heuristics and knowledge is proposed. In [10] a system was developed for extracting information from texts containing scientific citations; they consider a hybrid approach based on automatic learning, which combines text classification techniques with the Hidden Markov Model (HMM).

III. REFERENCE EXTRACTION STRATEGY

Similarly to references [2][7], the strategy used in this paper, for extracting bibliographic references, makes use of regular expressions. Besides that, classification techniques

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.12

such as the K-NN algorithm, the Naïve Bayes algorithm, Similarity of Cosine and Euclidian Distance are also employed here.

The extraction process receives as input a PDF file generated from an original digital document (PDF-text file) and produces as output the set of all references present in the paper, after the text "Reference" is first identified.

For the extraction process two strategies are combined to yield better results. The start of both strategies encompasses the following steps:

- Apply the pdfBox system [8] for reference extraction to PDF documents (papers) generating text format;
- Identify the word "Reference" in the text, considering several other conditions, that try to guarantee that this word is the one representing the reference section, such as:
  - Search for the word "Reference" both in Portuguese and in English, considering singular and plural forms;
  - Create a pattern for the word identified for Reference, by changing the identified combination to the standardized text "reference";
  - Identify if this word is in the beginning of the sentence, or if it is together with other words (like "bibliographic references").
- Identify the pattern "[number]", or sentences that have a quotation mark ("") and also ends with a number in year format.

For each reference, the information must be extracted, and each element classified as: title, author, place, year, or other information. The two proposed strategies are further detailed in next subsections.

A. *First strategy – Use of Regular Expressions*

The first strategy is based on Regular Expressions (RE). The experiments show that when the title is quoted, the other characteristics are normally identified with success. Figure 1 illustrates two styles of references in the same document: in the first one the title appears between quotation marks. The process is:

- First there is a search for a title between quotation marks. Then, the name of the author(s) is composed by the words that precede the identified title. The words that follow the title are classified as "other information", and from this group the year and the place will be obtained;
- As the author field generally has more than one author, this field must be split into: name and surname.

[4] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, and D. Veitch, "The multiscale nature of network traffic: discovery, analysis and modelling," IEEE Signal Processing Mag., vol. 19, pp. 28-46, May 2002.

[5] NORROS, I. A storage model with self-similar inputs. Queueing Systems, v.16, p.387- 396, 1994..

Figure 1. Example of Reference.

Looking at Figure 1, one may observe that besides the title appearing between quotation marks in the first reference and without them in the second, the format used for presenting the names of the authors is completely inconsistent. In the first reference ([4]) the initials appear in front of the family name of the author that is typed with the first letter in upper-case and the others in lowercase. In the second reference entry ([5]) the author surname appears first all in block capitals, followed by "comma" and the author initials. Notice that as an extra complicating factor, the last entry ([5]) has a typo and ends with two dots as ", 1994..".

Figure 2 exemplifies a case of failure in identifying the title of references using regular expressions: it is not between quotation marks; the author and the title are separated by comma; and in the title also includes a comma, but in this case, the comma does not indicate the end of the title.

[18] D. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.

Figure 2. Sample of Fail to Identify Reference with RE.

B. *Second Strategy - Automatic Classification*

The second strategy is based on Automatic Classification. For this strategy the references are divided into fragments, considering either the end point, or comma, or semicolon. For each fragment the user, in the training phase, classifies as Author, Title or Other. The algorithm analyzes the fragment and fills the vector with 24 features. The proposed vector is based on reference [10]. The vector has fields such as year, number, punctuation elements (".", ";", ",", "-" etc.).

Table 1 presents details about the addressed features and how the proposed vector details ranks for each element (title, author and other information), after training with 98 bibliographic references. There are features specific to a particular category, such as the element 2 (year) 98.63% that appears in the category "other information", and the element 12 (*et al.*) that is 100% in the "author".

The last two lines in Table 1, fragment length and position of the fragment, are used to indicate the size and position of the fragment, respectively, and this may help the classification. The last column of the table indicates the total fragments of each field.

The automatic classification is composed of two phases: training and test. For the training phase, the user fills the classifications of each fragment. For the test phase three strategies were applied: K-NN algorithm with Euclidean distance, K-NN algorithm with Cosine similarity and Naïve Bayes algorithm.

IV. EXTRACTION TOOL

This section describes the extraction system, which gives support to the proposal. The system has a specific interface for the **Training Phase**, which guides the following steps:

- The user inputs the references to be analyzed;

- The system partitions each supplied reference and find the punctuation marks: point, comma and semicolon;
- Each partition or fragment is analyzed by the system and fills in the corresponding vector of characteristics;
- Then, the user classifies each fragment, selecting its label as: title, author or other information;
- Finally, the system analyzes the training base generating general vectors with the measures obtained for each classification: title, author, and other information.

TABLE I. Vector of Features

Seq.	Fragment	Title	Author	Others	Total
		%	%	%	
1	Number	2,33	0,00	97,67	129
2	Year	1,37	0,00	98,63	73
3	Comma	19,60	26,40	54,00	250
4	Semicolon	50,00	33,33	16,67	6
5	Point	6,69	48,02	45,29	329
6	Dash or Two point	35,94	3,13	60,94	64
7	Ordinal Numbers	0,00	0,00	100,00	9
8	Word: ed., eds., editado	0,00	0,00	100,00	9
9	Word: p., pp., pag., pagina etc.	0,00	0,00	100,00	8
10	Word: v., vol., volume	0,00	0,00	100,00	7
11	Word: n., num., no., numero, número	33,33	0,00	66,67	3
12	Word: et al	0,00	100,00	0,00	1
13	word: and	40,00	33,33	26,67	30
14	Month	20,00	15,00	65,00	20
15	Country	0,00	0,00	100,00	1
16	Word: conference, symposium etc.	0,00	0,00	100,00	5
17	Word: available, doc, ps etc.	0,00	0,00	0,00	0
18	Preposition	81,63	0,00	18,37	49
19	Article	71,43	0,00	28,57	7
20	All uppercase	0,00	56,20	43,80	242
21	All lowercase	2,19	0,00	97,81	137
22	All capitalized	3,21	49,54	47,25	436
23	Fragment Length	-	-	-	-
24	Position of the fragment	-	-	-	-

Figure 3 shows the training phase and the accuracy obtained from the process.

The test phase is similar to the training phase, but the classification elements are automatically chosen by the system, and the vectors values are calculated. For each fragment, the system compares the calculated values with the General Vectors obtained. For the comparison phase three strategies are adopted.

The first one to be adopted is the K-NN algorithm with Euclidean distance, which is used to verify the similarity between the vector of fragments that is being considered with the general vector. The same is done using the K-NN algorithm with Cosine Similarity and finally the Naïve Bayes algorithm was tested.

After making some case studies using the 3 approaches, it was observed that K-NN algorithm with Euclidean Distance and Cosine Similarity present equivalent performance, that is, both achieved improvement in title extraction, but part of the authors still remain classified as "other information", which is not satisfactory. Tests using the Naïve Bayes algorithm, which considers the likelihood preceding element, showed better results. Figure 4 shows the Test Phase in block diagram.

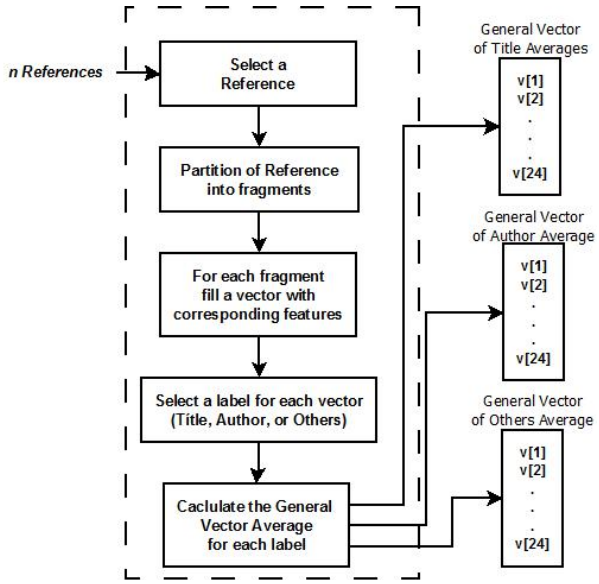

Figure 3. Training phase and extraction media.

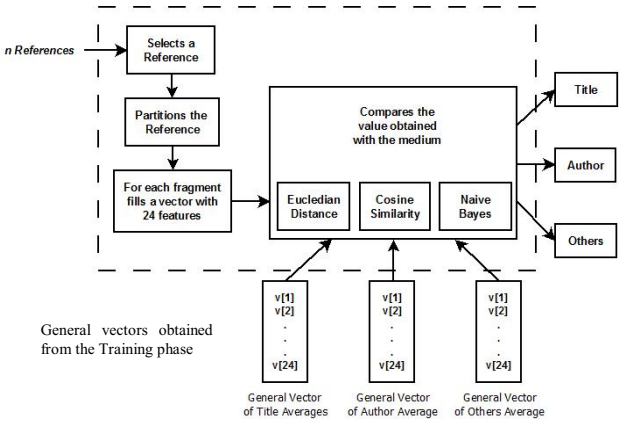

Figure 4. Test phase.

Figure 5 shows the training interface. The items are:

- 1 – indicates all references to be treated;
- 2 – the specific reference that is being treated;
- 3 – the element that is being considered at present, in this case the twelfth;
- 4 – which is the fragment;
- 5 – the classification informed by the system user, in this example the "year";
- 6 – the vector composed of 24 characteristics about the fragment being treated;
- 7 – buttons.

Figure 6 shows the test interface. Observe that, the interface is similar to the training interface; the difference is in the "result" fields. The items are:

- 1 – text that was classified as "title";
- 2 – text classified as "author";
- 3 – text classified as "other", and in this it possible extracted place and year with regular expression;

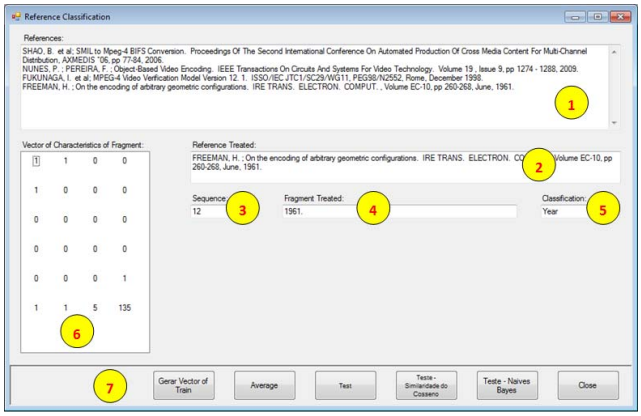

Figure 5. Training Interface.

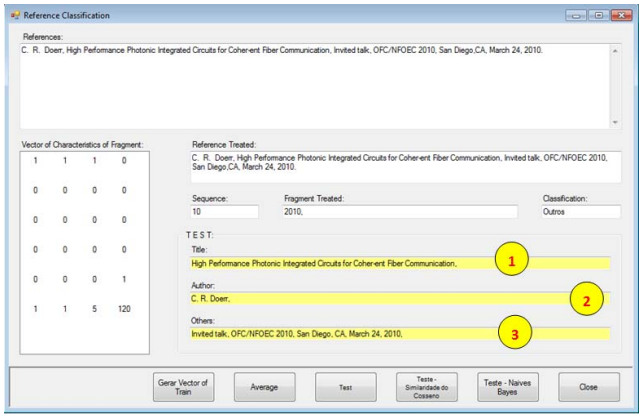

Figure 6. Test Interface.

Figure 7 shows an interface, from the Live Memory platform, where one or more articles can have their references extracted and classified. In that interface, the strategies with regular expression and the classifications (K-NN or Naives Bayes algorithms) were applied. For extracting the "place" and the "year" information, the system use regular expressions in "other information". Figure 7 is enumerated, indicating:

- Item 1 – the list of all references to be processed;
- Item 2 – all the "titles" found in the references of the selected article;
- Item 3 – all the "authors" found in the references of the selected article;
- Item 4 – all the "places" of publication found in the references of the selected article;
- Item 5 – all the "other fields" found in the references of the selected article;
- Item 6 – all the "years" found;
- Item 7 – the available buttons for cancelling the process or saving the extracted information in the database.

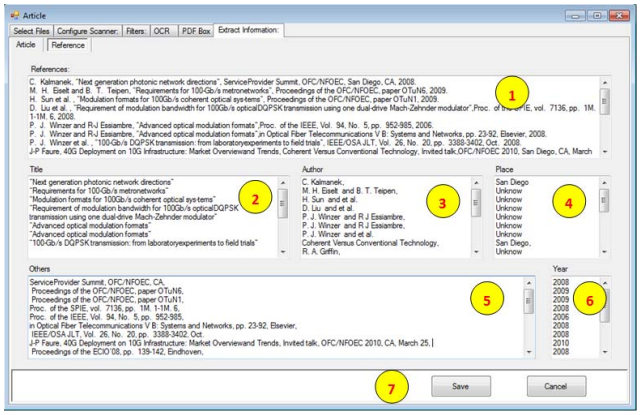

Figure 7. Article Interface: resulting table of the References.

V. RESULTS

In this section we will detail a specific study made to validate the proposal made in this paper.

A. Training

In the Training Phase, 16 papers were used, and from those 98 references were extracted and partitioned in 892 fragments, of which 363 fragments were classified as "titles", 417 as "authors" and 112 fragments were classified as "others".

The feature vector, described in Table 1, shows how these fragments were classified. One may observe that there are some cases where the feature is specific of a field, such as the "year", this is practically only given for the field "others".

If the title is in the database, the "other" fields are verified there also, eliminating the possibility of redundancy in the stored data, considering that there are references which have difference only in the year of publication.

B. Results

For validating the reference extraction scheme proposed here, the SBrT database was used [9], specifically the papers published in 2010, the papers are in English, because the event was the ITS'2010 - IEEE/SBrT International Telecommunications Symposium. Papers were in text editable PDF format.

At first, the focus was in the identification of references, and the strategy was to use only 10 papers from each year, which include 186 bibliographic references.

When the regular expressions were applied to the 10 papers, the following results were obtained:

- 117 “titles” were correctly identified, representing 62,90% of the existing tiles;
- 101 “authors” were identified, that is, 54,30% of the total;
- 170 “years” were identified, that is, 91,40% of the total;
- 122 were identified as “others”, and the accuracy rate was 65.59% of the total.

Table 2 shows the results when regular expressions are applied together with the Naïve Bayes algorithms. The classification of each reference element was:

- For “titles”, 138 were correctly identified, yielding 74,19% accuracy;
- For “authors”, the system correctly identified 114 instances, providing an accuracy of 61,29%;
- For “years”, 173 were correctly identified, that is 93,01%;
- For the “others”, 128 were identified corresponding to 68,82% accuracy.

TABLE II. Results

Field	Regular Expression (RE)		Naive Bayes and RE	
	Accuracy	(%)	Accuracy	(%)
Title	117	62,90	138	74,19
Author	101	54,30	114	61,29
Year	170	91,40	173	93,01
Others	122	65,59	128	68,82

Figure 8 also presents the same results, but in a graphic way. From the graphic it is clear that the one where the line is marked with an arrow represents a better solution.

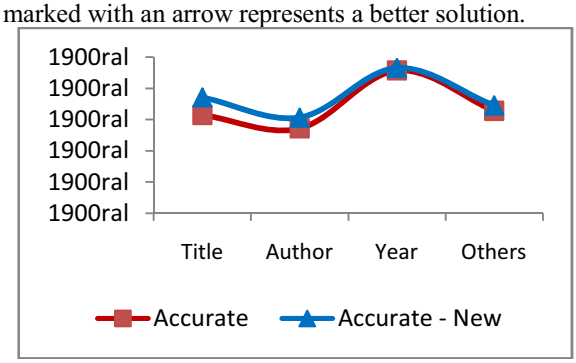

Figure 8. Graphic Results.

VI.CONCLUSIONS

The problem of reference identification and extraction from digital documents is not easy, despite of the existence of many proposals in the literature. In this paper, a specific process for reference extraction and classification of its sub fragments was detailed and tested; an implemented tool was used and also “real world” data, from existing conference documentation. Several algorithms and techniques were analyzed and tested, but the one that presented the best results were regular expressions and Naïve Bayes algorithms. Regular expressions work better if the identified elements in the references have a regular format, such as year; the result is not as good when it is required the identification of title or authors as they generally have no regular format. The integration of both strategies yields better results.

Despite of the process being tested using the support of the LiveMemory platform, the proposed process can be used independently of it, even the extraction tool presented.

ACKNOWLEDGMENT

This work was partially supported by CNPq-Brazil to whom the authors are grateful.

REFERENCES

[1] Abbyy FineReader Home Page. http://finereader.abbyy.com/.
[2] Álvarez, Alberto Cáceres. Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem.USP; 2007. Available at: http://www.teses.usp.br/teses/disponiveis /55/55134/tde-21062007-144352/.
[3] Bader Aljaber; Nicola Stokes; James Bailey; Jian Pei. “Document clustering of scientific texts using citation contexts,” Information Retrieval, V.13, N.2, 101-131, DOI: 10.1007/s10791-009-9108-x, 2009.
[4] Constans, Pere. “A Simple Extraction Procedure for Bibliographical Author Field,” Word Journal OF The International Linguistic Association, February, 2009, Available at http://arxiv.org/abs/0902.0755.
[5] Gupta, D.; Morris, B.; Catapano, T.; Sautter, G. “A New Approach towards Bibliographic Reference Identification, Parsing and Inline Citation Matching,” In Proceedings of IC3. 2009, 93-102.
[6] Hua Yang; Norikazu Onda; Massaki Kashimura; Shinji Ozawa. Extraction of bibliography information based on image of book cover. In Proceedings of the 10th International Conference on Image Analysis and Processing IEEE Computer Society Washington, DC, USA, 1999.
[7] Ohta, M., Yakushi, T, Takasu, A. “Bibliographic Element Extraction from Scanned Documents Using Conditional Random Fields” In Proceedings of ICDIM, 2008, 99-104.
[8] PDF-Box Home Page. Extracted from http://www.pdfbox.org, March 21 2011.
[9] R. D. Lins, G. Torreão, G. F. P. e Silva. Content Recognition and Indexing in the LiveMemory Platform GREC 2009. Springer Verlag. LNCS 6020. p.220-230, 2010.
[10] Silva, Eduardo Fraga do Amaral e. Um sistema para extração de informação em referências bibliográficas baseado em aprendizagem de máquina CIn-UFPE; 2004.
[11] Tesseract Home Page. Extracted from http://code.google.com/p/tesseract-ocr/downloads/list, June 22 2011.
[12] Atsuhiro Takasu, “Bibliographic Attribute Extraction from Erroneous References Based on a Statistical Model,” jcdl, pp.49, Third ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'03), 2003.