
Unsupervised Free-Text Processing and Structuring in Digital Archives

Alexander Ivanyukovich and Maurizio Marchese

Department of Information and Communication Technology, University of Trento, via Sommarive, 14, 38050 Povo (TN), ITALY

Automatic information extraction from unstructured sources is known to be a challenging task. It involves a number of sequential stages that transforms initial raw data to the source of knowledge in some field. In this paper we concentrate on the problem of the unsupervised free-text processing and structuring – 1st level of information extraction in the information pyramid we further introduce in this paper. We propose a novel approach that allows accurate text manipulation in order to detect coarse-grain structural skeleton within scientific digital archives. In contrast to other research initiatives in this area our approach relies on the information present in the processing text, does not depend on any external information sources (databases, dictionaries, semantic networks, etc) and is truly language-independent. Finally we discuss the evaluation methodology for the results obtained using our method and propose further research directions in the field.

Keywords: Information Retrieval, Text Recognition and Processing, Digital Libraries

1 INTRODUCTION

The exponential increase of information that current information retrieval and processing systems need to handle, requires support for automation of the majority of human-dependent parts of the process. We have classified these parts into five levels of the resulting information pyramid: text recognition (level-0), pre-processing and structuring (1st-level), metadata extraction and analysis (2nd-level), summarization (3rd-level), clustering (4th-level) and classification (5th-level). While the general case is hard to be completely automated, there is a visible progress in specific domains like scientific publication archives[1] and web-sites annotation in systems for Semantic Web[2]. Many of the most popular information retrieval systems (for instance Google[8], CiteSeer.IST[1], Vivisimo¹, Clusty², etc) cover subsets of the above information processing stack.

The main bottleneck of the levels transition is between the required quality and level of automation of the process. The problem roots in the natural language formalization research domain and there is no solution for generic case at the moment. In our overall research activity we are aiming to investigate a completely automated information extraction stack while pushing the domain coverage on the possibly highest level. We believe that accuracy of automatic information extraction at the higher level of the pyramid can be achieved only by addressing the high quality at the lower levels. The pyramid starts from the text recognition step that includes text extraction from specific text-based formats (DOC, RTF, etc) as well as text identification using Optical Character Recognition (OCR) techniques in case of non-textual formats. Current work addresses 1st level of the information extraction stack: text pre-processing and structuring. In particular, we focus on a language-neutral approach for automatic processing of free text contributing to the previous broad goal.

The main contribution of this paper is a novel approach to unsupervised text pre-processing and structuring that contributes to the high-quality information extraction on the next levels. The approach is language-neutral, i.e. it can be applied to any spoken or artificial language that does not violate general assumptions summarized further. It does not rely on any external information sources (dictionaries, knowledge-bases, oracles, etc) and is solely based on the information available within the processing text and its context. Further in this paper we discuss in detail three components of our approach, namely text flow recognition, hyphens correction and collateral elements detection (for instance index, table of content, etc).

The problem of unsupervised free text pre-processing is under study for a long time in the context of metadata extraction. This include duplicate and near-duplicate texts identification in large-scale information retrieval and processing systems[7,8], dictionary-based text structure analysis and partial processing [1], application of statistical models like Hidden Markov Model (HMM)[10] and Dual and Variable-length output Hidden Markov Model (DVHMM)[11] for sequences prediction (punctuation, context breach, etc) as well as Support-Vector Machines (SVMs) and other Machine Learning methods application to the text pre-processing and structuring. The outlined approaches are different by their nature, but all of them need some

¹ <http://www.vivisimo.com>

² <http://www.clusty.org>

manually prepared information source (training set, dictionary, etc) and are mostly language-dependent.

The remainder of the paper is organized as follows. In Section 2 we describe in detail the proposed approach for 1st-level text processing: in particular we describe a three step procedure that is based on the statistical methods and a-priori domain-specific knowledge. In Section 3 we describe how the proposed approach has been applied to a large set of documents (ca. 120 K) and we discuss possible methodologies for its evaluation and comparison with the state of the art in the domain. Section 4 summarizes the results and discusses our future work.

2 THE APPROACH

The approach we propose for unsupervised free text pre-processing and structuring consists of three major steps, namely (1) text flow recognition, (2) hyphens correction and (3) collateral elements identification. The methods were originally introduced to solve the discussed problem in the large scientific articles corpora, however, the evaluation discovered that the same techniques can be used in some other domains as well. In our experiments we utilized a subset of ca. 120K scientific documents from the CiteSeer.IST[1] collection. The set has some specific features that can be missing in the generic free text collections, for instance guaranteed existence of the pages in each text in the explicit form, because of the texts acquisition method (conversion from PDF and PostScript sources).

There are a number of goals we aimed for the 1st-level in our information extraction pyramid. Text pre-processing contribute to the more accurate textual information acquisition, i.e. correctly identified text flow (pages ordering), removal of the repeated elements that do not contribute to the semantic and structural content (headers and footers) and removal of text delimiters inside structural elements (footnotes and page numbers inside single reference, hyphens in the authors' names and titles in references, etc). Text structuring contributes to the correct identification of the major structural elements within a text, i.e. Introduction section and reference to the Introduction section in a Table of content section of the article should be correctly distinguished and handled appropriately. The presented approach aims to be independent on the external information sources through identification of the needed information in the processing documents collection. In the subsequent sections we describe each processing step in detail.

2.1 Text Flow Recognition

The problem of the text flow recognition in general goes to the language formalization problem and at present it does not have a general solution. In our approach we focus on the specific problem of global text flow recognition, assuming that local text flow is already in the correct order. This situation exactly describes the pagination case of the document when, within the page, text flow is already correct, while the pages themselves can be mixed up (in case of our dataset pages can be either in normal or in reverse order).

Numbers succession method is the most easy and straightforward method for text flow recognition. It is based on the analysis of the number sequences at the top and bottom of the pages. Complex situations include existence of tens of sequences across the text, discontinuous sequences (chunks) and competing sequences. The decision should be made through calculation of the longest number sequence, possibly involving some heuristics. The method works well in most cases for well-formatted articles with explicit page numbering (see **Figure 1**). However, it collapses in partial presence or absence of the pages numbering and presence of other digits sequences not related to the text flow. Example of a false positive situation for this case is presented on the **Figure 2**.

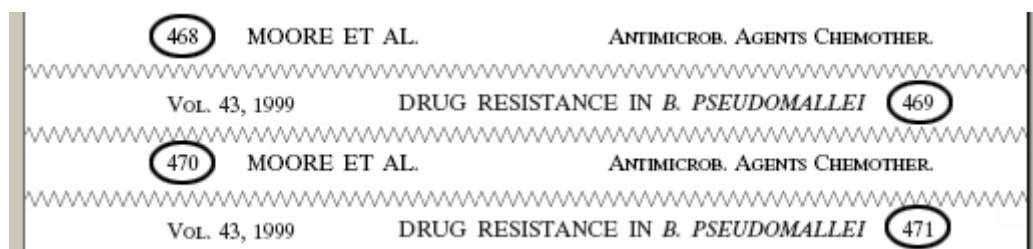


Fig. 1. Example of true positive situation for numbers succession method (headers)

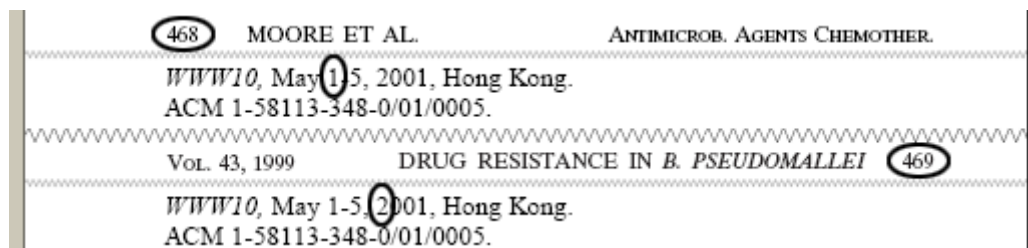


Fig. 2. Example of false positive situation for numbers succession method (header/footer combinations)

Hyphens concatenation method requires big text corpora for dictionary creation. It is based on the existence of situations when a word is divided between two different pages. Collation of the pairs for lexical coincidence allows detection of the natural text flow. There is no dependency on the text language, because the control dictionary is build from the current processing corpora. Absence of the explicit hyphen does not influence the method applicability, but rather on the calculations speed. The problematic situation for this method can be existence of the header or footer between checked word's parts. Example of the described problematic situation for this algorithm is presented on the **Figure 3**.

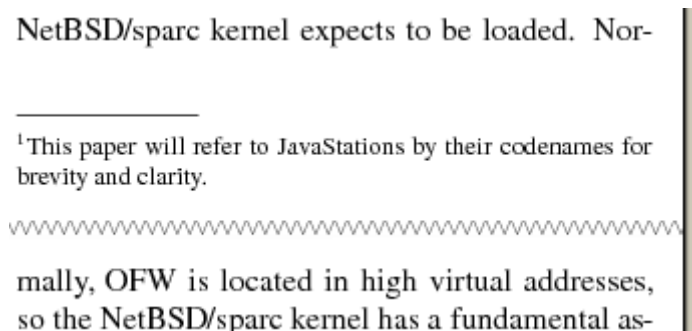


Fig. 3. Example of a problematic situation for hyphens concatenation method.

Text flow prediction allows eliminating the dependency on words division between pages as described in the previous method. In the contrary it involves prediction of the possible words combinations on the page borders using Hidden Markov Model (HMM) or Dynamic Bayesian Networks (DBNs) [9]. The application restrictions are similar to the ones described in the previous method.

Weighted structure method overcomes the limits of the presented methods but it is more complex and depends on the text formatting and size of the processing corpora. The method is based on structure elements identification using dictionary and their order comparison analysis. Dictionary creation involves identification of the structure candidates across the current corpora (specially formatted words and words combinations) and their frequency-based rating. The normal text flow is assumed to follow the same sequence as the statistical maximum sequence from the dictionary. Problematic situations for this method include existence of duplicates of the structure elements in the text body (for instance in the table of content and document index), but this can be solved using special heuristics.

2.2 Hyphens correction

Correction of hyphens allows correct semantic and structural elements identification and processing. Examples include: authors identification, normalization and comparison, titles comparison, and identification of the elements' borders. Normally the task described requires manually prepared dictionary in order to obtain high-quality results, introducing additional non-trivial task of text language identification. Here, to overcome these difficulties we propose using automatically prepared dictionary based on the existing texts corpora.

2.3 Collateral Elements Identification

Text structure detection is a non-trivial problem in a general case, requiring application of the Natural Language Processing (NLP) methods. We suggest addressing part of this task by eliminating collateral structural elements thus decreasing the probability of error during structural elements detection using other techniques. This contribution allows improving the quality of metadata extraction on the 2nd-level processing by eliminating possible false positive situations.

Based on the analysis of the processing texts corpora we were able to define two different collateral

elements' types: (1) structural elements (table of content, index, etc) and (2) non-structural elements (page header and footer). In the suggested method we address both types assuming the following manually verified properties to be valid within the corpora:

- Structural elements presentation on the page are more sparse than the rest of the text;
- Structural elements are normally represented by single semantic construct (sentence, word or word combination);
- Structural elements often have some denomination at the beginning;
- Structural elements normally have structure identifiers in a form of digits (or bullets) with order and optional indentation;
- Structural elements can be divided by non-structural elements which can raise a problem of correct elements borders identification;
- Non-structural elements are normally lexically agree within the punctuation through the whole text;
- Non-structural elements can repeat semantic constructs within the same page or subsequent number of pages (header "Introduction" on the page with chapter "Introduction").

Using statistical analysis of the words-per-line and its average over the whole text we are able to correctly identify most of the collateral structural elements within the test corpora. The remaining true negatives situations are handled using simple pattern-based Finite State Machine (FSM). Identification of the non-structural elements is accomplished using either set of patterns (as we did in our case) or involving mathematical models for punctuation prediction.

3 EVALUATION OF RESULTS

Since we are working with sets of hundreds of thousand of documents, manual quality evaluation for the whole collection is not feasible. This fact has been already reported in related works [4,5,6] and other methodologies were proposed, based on the specifics of each concrete dataset. Moreover it is difficult to obtain suitable dataset on which to compare the various methods. To overcome this problem we based our evaluation on a generic evaluation of the overall process – i.e. to obtain the same results using our methods precision on the described step should not be less than current precision in existing state-of-the-art systems. In particular, in our experiments we compute quality evaluation criteria from preliminary results (extracted metadata) obtained with the proposed approach with results obtained in CiteSeer.IST. In order to build a consistent reference dataset for our evaluation, we have processed results of the CiteSeer.IST Autonomous Citation Indexing (ACI) system, publicly available within its Open Archive Initiative (OAI). The results contain metadata for 570K documents collected within CiteSeer.IST project within the last several years. We have used this metadata collection as an "ideal" metadata. Here "ideal" means that this metadata was processed using external manually verified datasets (DBLP) and afterwards manually re-checked and corrected by users. Therefore, for our purpose, it is our "ideal" reference set.

We have processed list of URLs available within the "ideal" collection and were able to retrieve from the Internet around 120K of documents (other 450K documents have disappeared from their original location after they were collected by CiteSeer.IST - in 3-4 years time). In this subset of documents, we were able to process, correctly identify and extract metadata from more than 90K of documents. This set corresponds to our complete "ideal" collection used in our preliminary evaluation.

For our evaluation task, we have used standard quality measure criteria, namely: precision, recall, and F-measure. In Figure 3, we report the definition of the quality criteria, where:

- A is the number of true positive samples predicted positive; in our case A is the number of references that we have recognized in a document that is not more than the number of references in the corresponding document in the ideal set.
- B is the number of true positive samples predicted as negative; in our case it is the difference between number of references present in the ideal set and number of reference in corresponding document processed with our approach - but not less than zero.
- C is the number of true negative samples predicted as positive; in our case it is the difference between results obtained with our approach and the ideal set - but not less than zero.

$$Precision = \frac{A}{A + C}$$

$$Recall = \frac{A}{A + B}$$

$$F_{measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Fig. 3. Evaluation criteria: Precision, Recall and F-measure.

$$Precision = 87,75\%$$

$$Recall = 88,52\%$$

$$F_{measure} = 88,14\%$$

Fig. 4. Evaluation results

In Figure 4, we report the results of our evaluation tests. These preliminary results show that our approach is capable of achieving the quality level that is comparable with the one present in the ideal set but without usage of external manually verified datasets and without human supervision (which is the case for the ideal set).

4 CONCLUSION

In this paper we have presented a novel approach to unsupervised text pre-processing and structuring that contributes to the high-quality information mining on the next levels of information extraction pyramid. In the contrast to existing methods our approach does not rely on the external information sources and is language neutral. Further evolution of the approach is possible and according to the evaluation results obtained will cover integration of the presented approach with methods that use external knowledge-bases and other sources.

ACKNOWLEDGMENT

We acknowledge C. Lee Giles for useful comments and advice during initial brainstorming on the system architecture as well as Fausto Giunchiglia for his advice and continuous support to the research project.

REFERENCES

- [1] Giles C.L., Bollacker K., Lawrence S. CiteSeer: An Automatic Citation Indexing System. Digital Libraries 98: Third ACM Conf. on Digital Libraries, ACM Press, New York, 89-98, 1998.
- [2] Erdmann M., Maedche A., Schnurr H.-P., Staab S. From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, Luxembourg, 2000.
- [3] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, ACM Press, 34(1):1-47, 2002.
- [4] Vaclav Petricek, Ingemar J. Cox, Hui Han, Isaac G. Council, and C. Lee Giles. A comparison of on-line computer science citation databases. In Proceedings of the European Conference on Digital Libraries, Lecture Notes in Computer Science. Springer, 2005.
- [5] Hui Han, C. Lee Giles, Eren Manavoglu, Hongyuan Zha, Zhenyue Zhang, and Edward A. Fox. Automatic document metadata extraction using support vector machines. In Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, pages 37-48. IEEE Computer Society, 2003.
- [6] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plaintext collections. In Proceedings of the 5th ACM International Conference on Digital Libraries. ACM, 2000.
- [7] Allan Heydon and Marc Najork. Mercator: A Scalable, Extensible Web Crawler, World Wide Web, 2(4):219-229, Dec. 1999.
- [8] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of the 7th International World Wide Web Conference, pages 107-117, Brisbane, Australia, April 1998. Elsevier Science.
- [9] Leonid Peshkin and Avi Pfeffer. Bayesian Information Extraction Network. In Proceedings of IJCAI 2003.
- [10] Silviu Cucerzan and David Yarowsky. Language independent, minimally supervised induction of lexical probabilities. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pages 270-277. Association for Computational Linguistics, 2000.
- [11] Atsuhiro Takasu. Bibliographic attribute extraction from erroneous references based on a statistical model. In Proceedings of the 2003 Joint Conference on Digital Libraries. IEEE, 2003.