INFORMATION
SYSTEMS IN
MANAGEMENT

Information Systems in Management (2014) Vol. 3 (4) 261–272

ADAPTIVE INFORMATION EXTRACTION FROM STRUCTURED TEXT DOCUMENTS

PIOTR OŻDŻYŃSKI, DANUTA ZAKRZEWSKA

Institute of Information Technology, Lodz University of Technology

Effective analysis of structured documents may decide on management information systems performance. In the paper, an adaptive method of information extraction from structured text documents is considered. We assume that documents belong to thematic groups and that required set of information may be determined ”apriori”. The knowledge of document structure allows to indicate blocks, where certain information is more probable to appear. As the result structured data, which can be further analysed are obtained. The proposed solution uses dictionaries and flexion analysis, and may be applied to Polish texts. The presented approach can be used for information extraction from official letters, information sheets and product specifications.

Keywords: Natural Language Processing, Information Extraction, Tagging, Named Entity Recognition

1. Introduction

Problem of information extraction consists in identifying specific information from a set of documents [1]. Kanya and Ravi [2] defined information extraction as the task of recognizing mentions of entities and their relationships in text documents. Accordingly classical information extraction tasks include Named Entity Recognition (NER), which addresses the problem of the identification and classification of predefined types of named entities. This kind of problems are very often connected with structured documents processing and analysis, which constitute very important part of management information systems functionalities, as many of documents used in organizations comprised of the structured ones.

In the paper we consider structured documents belonging to thematic groups, where predefined type of information is sought. We assume that document structures are "apriori" identified, what allows us to indicate blocks, where required information is expected to appear. In the proposed approach information extraction problem is treated as classification task, where instead of patterns sets of descriptive predicates are used.

The presented adaptive algorithm of information extraction aims at specifying structured data, convenient for further analysis. The algorithm uses dictionaries and flexion analysis and can be applied to Polish language documents.

The paper is organized as follows. In the next section relevant research concerning named entity recognition methods and their applications is presented. Then the proposed methodology is depicted and the system architecture is described. The following section is devoted to two case studies of documents connected to internet shopping cards. In Section 6 experiments concerning considered case studies are presented and obtained results are discussed. Finally some concluding results and future research are depicted.

2. Related work

Task of information extraction from structured documents was investigated by many researchers. There exist different approaches for solving this problem. However most of them concern English or Chinese documents. In the paper [3], authors used keyword matching for knowledge extraction. For this purpose they built a knowledge-oriented Web page automatic acquisition system (AKAS2WP). Cvitaš considered methods of information extraction based on relation detection [4]. His methodology concerns finding out already defined relations between formerly mentioned entities. Relation extraction aims at exploiting the text and transforming it into structured source. Effectiveness of machine learning techniques is presented taking into account supervised, unsupervised as well as weakly unsupervised approaches. Special attention was done to the last one: the author developed co-learning method and compared its effectiveness to supervised techniques.

Conditional Random Fields (CRF) approach plays a significant role among information extraction techniques. In the paper [5] some improvements of CRF were proposed. Authors include the domain ontology features to CRF's model, to enable using of semantic information. What is more, thanks to building compound features, authors could use more rich overlapping features, by combining internal and external ones. Kosala et al. [1] consider the problem of extracting specific information from a collection of structured documents. They develop a wrapper induction method that utilizes the tree structure of the documents. Tree automata are used as wrappers. The authors compared the proposed approach to string automata. They stated that exploiting the tree structure augmented the performance of information extraction. As the conclusion advantages and disadvantages of the proposed method were indicated.

Researchers examined many NER techniques. In [6], joint approach of NER and relation identification to information extraction from Web documents was considered. Authors concluded that using of relation extraction helps in improving the performance of NER task. Semi supervised NER method was examined in [7]. The proposed technique was based on internal and external pattern fusion. NER performance was improved by using soft-matching and bootstrapping. Authors showed that the proposed approach gives better results comparing to traditional NER approaches. Zhu proposed an adaptive NER framework to extract information from the Web [8]. He integrated different techniques to recognize entities of various types on various domains. In the presented approach he discovered a hierarchy from Web link structures by taking into account links between pages and page content. The domain hierarchy allowed to select effectively domain vocabulary and patterns, which were formalized as association rules. Kanya and Ravi [2] presented and compared several NER methods. They showed that machine learning techniques are promising tools for solving information extraction tasks. They limited their investigations to three methods: Hidden Markov Models, Transformation-Based Learning and Support Vector Machines. They stated that, the last model outperformed the others in the cases they considered. Todorović et al. applied probabilistic generative Hidden Markov Model combined with grammar based component to extract information from unstructured documents [9]. They used context of surrounding words to obtain more accurate results.

Named entity recognition approach was used not only for supporting management information systems. Many authors investigated application of the method for documents from biomedical area (see [10], [11], [12] for example).

3. Methodology

The problem of information extraction from text documents may be considered as the classification task, which aims at finding out an approximation of an objective function of a form:

$$\Phi\colon D \times C \to \{T, F\}, \tag{1}$$

where C is the set of predefined categories (classes) and D is the set of documents.

Function defined by (1) is called a classifier and for each pair *document – class*, determines a logic value *true* if the document belongs to the certain class from C and *false* if it does not [13]. In the information extraction case document class is equivalent to the information content of the document.

There exist many document classification methods [14]. Most of them are based on machine learning techniques, such as kNN, Naïve Bayes [2] as well as mining for frequent sequences or emerging patterns [15]. However for their effective use sufficiently big training set is required. In the presented method, as counterpart of emerging patterns, sets of descriptive predicates are proposed. Thus a sentence function describing a single predicate takes the form:

$$f(doc, seq, tag) \to \{T, F\}, \tag{2}$$

where *tag* is considered as integral text fragment together with its role in the document, *seq* means the tag sequence representing structurally separated document block containing argument *tag* and finally *doc* means an analysed document in the form of block segments comprising argument *seq*.

Each document can be classified only if all the conditions of the class membership are fulfilled. However the conditions are constructed in a way that allows to include alternative and negation operations, what enables flexible pattern defining. The presented function is valued for every tag in a document context. If the predicate is fulfilled, the document is marked by class label.

4. System architecture

The proposed system aims at finding previously defined information in the input document set. The system works on preprocessed documents, where text is divided into blocks by lines including their names, in the following form:

```
#[p]
First block content
#[p]
Next block content.
```

For different data sources document formats may differ significantly and each of them requires disparate preprocessing procedures. For example completely different approach is necessary for preparation of *html* and *doc* documents. As considering of different document formats exceeds the scope of the current research, the presented system works on sets of previously prepared documents.

During the first stage, text is divided into indivisible fragments, for which tags connected with their meanings are assigned. We assume that each indivisible fragment consists of characters of a single type such as letters or digits or the other characters like punctuation marks. To assure univocal character nature the set of tags allocated to texts is determined in advance.

Set of conditions, used for information extraction, is defined in the way which enables to determine requested sequence unequivocally taking into account the smallest number of conditions the possible. To avoid problems generated by language flexion the system uses regular expressions, which allow to define requested words in different grammatical forms. Predicates are constructed in a way, which enables finding easily respective information. Such characteristics is obtained by using similarity of text document structures. Predicates are modified during the consecutive iterations. Dictionaries are supplemented during analysis process, thus the system can be easily adapted to considered documents. The whole system architecture is presented on Fig. 1.

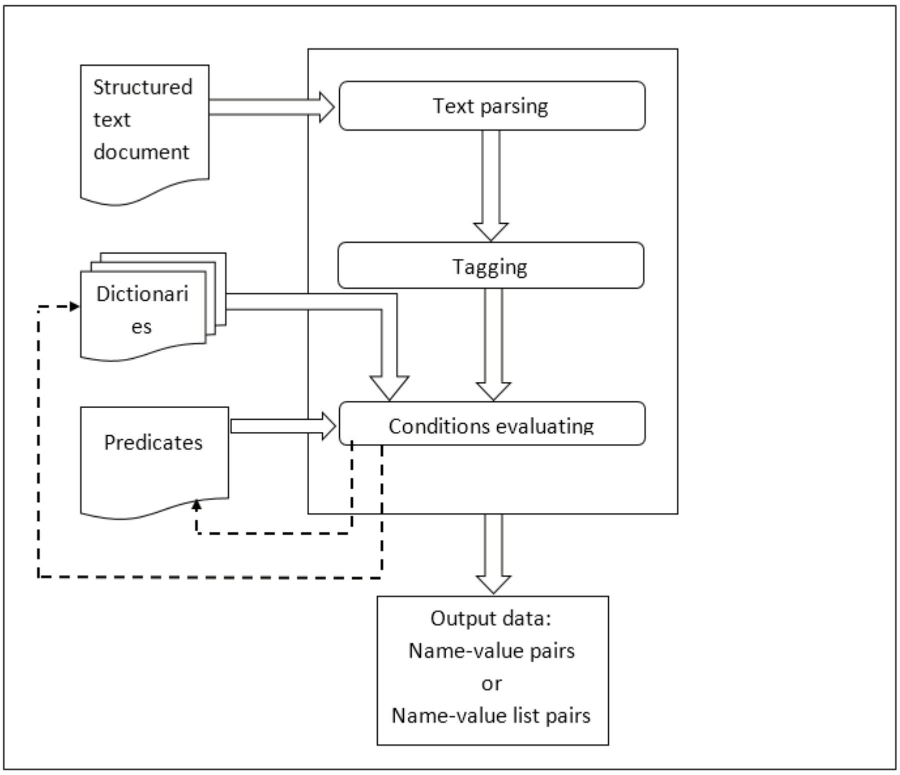

Figure 1. The system architecture

5. Case studies

For the experiment purpose, we consider two case studies of product cards, which can be found on internet shops web sites. The first case concerns descriptions of sportive gloves for mixed martial arts (MMA). There are three main features, which distinguish the gloves: *type* (purpose), *size* and *material type*. All of them will be the matter of inquiries in the information extraction process.

In the second case documents describing bikes were considered. As previously, three attributes: *name*, *frame size* and *brake type* were taken into account.

For the purpose of the both document sets the following conditions were considered:

- *tag* – the condition is fulfilled if it is consisted with the required one,
- *regexp* – the value fits to the given regular expression,
- *sequence* – indicated tag sequence (usually several preceding words) fulfils required conditions (all or one),
- *dict* – tag value is in the dictionary.

As conditions can be nested we propose to use XML language for predicates descriptions. An example of type of gloves definitions is presented on Fig. 2.

The condition defined on Fig. 2 can be interpreted as follows: document will be labelled by the text "rękawice" (value taken from the suitable tag) if the text fragment tagged by WORD (consisting only of letters) is found. What is more one of two words preceding this tag will be the word "rękawice". The last condition means that the value can be found in the dictionary containing glove types. If all the conditions are fulfilled then the suitable tag will be the word denoting the glove type and the class label will be created from the pattern of attribute value.

```
<output type=″name″ value=″Rękawice {this}″ hint=″FIRST″>
<condition tag=″WORD″ />
<condition type =″ANY_MATCHES″ subject=″SEQUENCE″
from=″-2″ to=″0″>
<condition regexp=″(Rękawice)|(rękawice)|(RĘKAWICE)″/>
</condition>
<condition dict=″glove-types″ />
</output>
```

Figure 2. Definition of type of gloves

6. Experiment results and discussion

Experiments aimed at checking, how the proposed approach works for the two considered case studies. In the first step the set of 114 partially structured documents concerning sportive gloves were considered. The experiments were carried out assuming that effectiveness of 100% in the information extraction process is required, taking into account previously prepared predicates.

For the predicate determining that the type of gloves is contained in the dictionary and occurs after the word "rękawice", 104 documents were labelled. The remaining 10 documents were classified after adding the second predicate, which based the definition on the sequence of words as the *type name* consisted of several words.

The next required information concerns the *material type*. The first predicate classified 85 documents as "skóra naturalna". Predicate consisted of the condition concerning the presence of a word beginning by letters "skór", which occur in the neighbourhood of the words "naturalna" or "bydlęca". The following 21 documents fit to predicate describing "skóra syntetyczna" or "ekologiczna" and were labelled by "skóra syntetyczna". The last 8 documents required 3 predicates determining materials "XDsyntex", "DX" and "materiał skóropodobny".

Finally, glove *sizes* were considered. In this case one predicate was sufficient to classify 20 documents, two predicates were necessary to label 61 documents, 12 documents required 3 predicates. The remaining 21 descriptions did not contain the size information.

Thus, the system classified all the documents according to glove *type*, *material* and *size* by using 3 predicates at the most.

During the second part of experiments, bike product cards were taken into account. Two datasets, A and B, each consisting of 100 documents from 10 different sellers were examined. Each seller was represented by 10 cards. The set A was applied to build predicates, which were further used to find out the required information from the set B. Investigations concerned *bike name*, *frame size* and type of *brakes*. Predicates were created iteratively.

In the case of the attribute *bike name*, 53 documents were necessary for creating 7 predicates to label all the documents. Obtaining the final results required 41 iterations. In the case of the *frame size*, analysis of 8 documents and 3 predicates were necessary to label the number of 57 documents, the all ones for which *frame size* was determined. Final effects were obtained in 6 iterations. The attribute *brake type* was used in 85 documents. This number of documents was labelled in 9 iterations. 7 predicates built by analysis of 9 documents was necessary to label all the required documents.

Effects of document labelling for all the attributes and different number of documents, are shown in Table 1. The first two columns present attribute and the iteration number, the third column contains the number of considered documents, the next one shows the number of predicates used to label the documents, which number is presented in the last column. All values are presented in a cumulative way.

Table 1. Document labelling for bike product

Attribute	No iter.	Documents analysed	Predicates	Labelled documents
Name	1	2	1	10
	2	12	4	29
	11	21	4	42
	17	27	5	56
	19	30	6	59
	41	53	7	100
Frame size	1	3	1	22
	4	6	2	52
	6	8	3	57
Brake type	1	1	1	18
	3	3	2	27
	4	4	3	33
	5	5	4	39
	6	6	5	66
	7	7	6	68
	9	9	7	85

After this part of experiments, for each attribute we obtain a set of predicates, which can be further used to extract information from the set B. Comparisons of effects obtained for both of the datasets for all the attributes are presented in Table 2. Columns contain numbers of documents labelled in the set A and B respectively.

Table 2. Results obtained for the both datasets

Attribute	Set A	Set B
Name	100	85
Frame size	57	52
Brakes	85	86

It is easy to notice, that the effects obtained for both of the datasets are similar. Big number of attribute values, in the case of *bike name* resulted in less number of documents labelled in the set B. In the case of *frame size* and *brake type*, numbers of attribute values are significantly smaller, thus dictionaries are limited and numbers of necessary iterations considerably diminish.

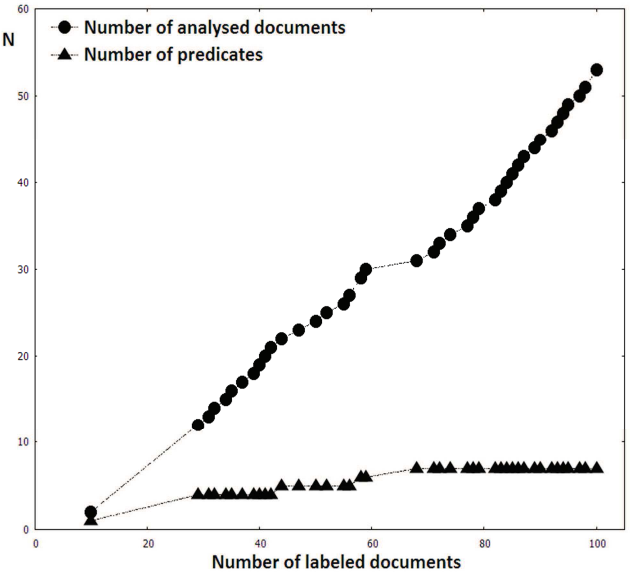

Figure 3. Documents and predicates for *bike name*

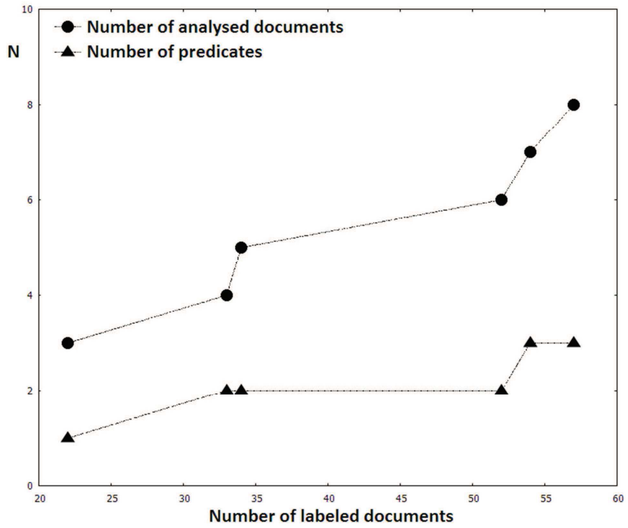

Figure 4. Documents and predicates for *frame size*

New predicates are not defined in each iteration. After analysis of consecutive documents, new names are added to dictionaries. Availability of dictionaries, before the process starts, would ensure the smaller number of iterations necessary to label all the documents. However creating dictionaries during analysis makes the system flexible and independent on input data.

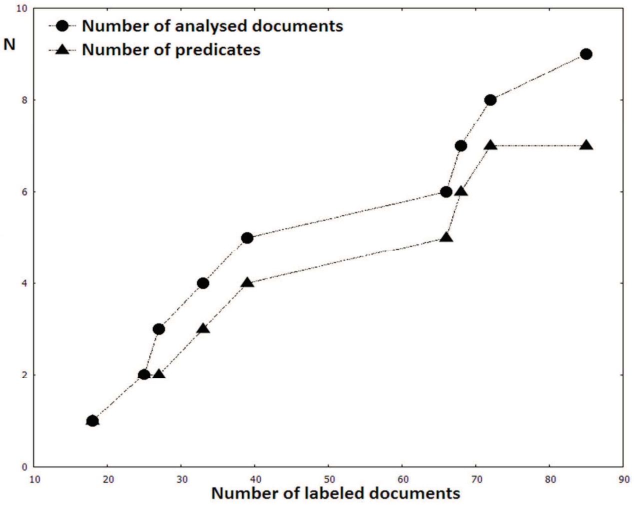

Figure 5. Documents and predicates for *brakes*

Finally, it is worth to notice that the numbers of required predicates do not exceed 7 for all the considered attributes and that they do not depend on the iteration numbers. Charts showing these dependencies for all the attributes are presented respectively on Figures 3, 4 and 5. In all the cases, axis OX shows numbers of labeled documents while axis OY presents iteration numbers.

7. Conclusion

In the paper the method of information extraction from structured documents of a certain thematic group is proposed. The method is based on sets of predicates, uses dictionaries and flexion analysis and can be applied for Polish language documents. The presented algorithm is evaluated by experiments carried out on the datasets of internet product cards. Obtained results showed that despite its simplicity the proposed method performs well and can be used during preprocessing phase of text analysis for information extraction concerning previously determined attributes.

The proposed approach enables to adapt the system to new documents, when the considered dataset is extended. While adding new documents, dictionaries are actualized and developed, predicates are modified or the new ones are created.

The considered method may be applied in internet shops, where several products are compared taking into account chosen features. Automation of parameter generation allows to analyze products from different suppliers, who do not have to adjust considered information to required format. Adding new groups of products is connected with constructing new predicates, which will ensure getting expected effectiveness.

Future research will consist in considering other applications of the proposed method, such as medical document analysis as well as optimization of predicates construction, which will enable improving classification performance.

REFERENCES

[1] Kosala L., Blockeel H., Bruynooghe M., Van den Bussche J. (2006) *Information Extraction from Structured Documents Using k-testable Tree Automaton Inference*, Data & Knowledge Engineering 58, 129-158.

[2] Kanya N., Ravi T. (2012) *Modeling and Techniques in Named Entity Recognition - An Information Extraction Task*, Third International Conference on Sustainable Energy and Intelligent Systems, Tamilnadu, India, 27-29 December.

[3] Zhu Junwu, Jiang Yi, Xu Yingying (2009) *Automatic Knowledge Acquire System Oriented to Web Pages*, Proc. of the 3rd International Conference on Intelligent Information Technology Application, 21-22 Nov., Yangzhou University Yangzhou, China, 487-490.

[4] Cvitaš A.(2011) *Relation Extraction from Text Documents*, Proc. of the 34th International Convention MIPRO 2011, May 23-27, Opatija, Croatia, 1565-1570.

[5] Fang Luo, Pei Fang, Qizhi Qiu, Han Xiao (2012) *Features Induction for Product Named Entity Recognition with CRFs*, Proc. of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design, 491-496.

[6] Xu Qiuyan, Li Fang (2011) *Joint Learning of Named Entity Recognition and Relation Extraction*, 2011 International Conference on Computer Science and Network Technology, 1978-1982.

[7] Cheng Ziguang, Zheng Dequan, Li Sheng (2013) *Multi-Pattern Fusion Based Semi-Supervised Name Entity Recognition*, Proc. Of the 2013 International Conference on Machine Learning and Cybernetics, Tianjin, 14-17 July, 45-49.

[8] Zhu Jianhan (2009) *An Adaptive Approach for Web Scale Named Entity Recognition*, 1st IEEE Symposium on Web Society 2009, 41-46.

[9] Todorović B.T., Rančić S.R., Marković I.M., Mulalić E.H., Ilić V.M. (2008) *Named Entity Recognition and Classification using Context Hidden Markov Model*, 9th Symposium on Neural Network Applications in Electrical Engineering, September 25-27.

[10] Chan Shing-Kit, Lam Wai (2007) *Efficient Methods for Biomedical Named Entity Recognition*, Proc. of the 7th IEEE International Conference on Bioinformatics & Bioengineering, Boston MA, October 14-17, 729-735.

[11] Liao Zhihua, Wu Hongguang (2012) *Biomedical Named Entity Recognition based on Skip-Chain CRFS*, 2012 International Conference on Industrial Control and Electronics Engineering, 1495-1498.

[12] Keretna S., Lim Ch. P., Creighton D. (2014) *A Hybrid Model for Named Entity Recognition Using Unstructured Medical Text*, Proc. of the 2014 9th International Conference on System of Systems Engineering, Adelaide Australia, June 9-13, 85-90.

[13] Debole F., Sebastiani F. (2005) *An analysis of the relative hardness of reuters-21578 subsets*, J. Am. Soc. Inf. Sci. Technol., 56/2005, 584-596.

[14] Sukanya M., Biruntha S. (2012) *Techniques on text mining*, Proc. of the IEEE Int. Conference on Advanced Communication Control and Computing Technologies, 269-271.

[15] Ożdżyński P. (2014) *Text document categorization based on word frequent sequence mining*, Information Systems Architecture and Technology, Contemporary Approaches to Design and Evaluation of Information Systems, Oficyna Wydawnicza Politechniki Wrocławskiej, 129-138.