INFORMATION SYSTEMS IN MANAGEMENT

Information Systems in Management (2014) Vol. 3 (4) 229−238

AUTOMATIC INDEXER FOR POLISH AGRICULTURAL TEXTS

WALDEMAR KARWOWSKI, PIOTR WRZECIONO

Department of Informatics, Warsaw University of Life Sciences - SGGW

Today, the majority of resources are available in digital forms to acquire information. We have to search through collections of documents. In this paper text indexing which can improve searching is described. Next, indexing tool, the Agrotagger, which is useful for documents in the field of agriculture, is presented. Two available versions of the Agrotagger are tested and discussed. The Agrotagger is useful only for the English language despite the fact that it uses multilingual thesaurus Agrovoc. Because of the Agrotagger is not useful for texts in Polish, it is important to create similar tool appropriate for the Polish language. The problems connected with extensive inflection in languages such as Polish language in the process of indexing were discussed. In the final part of the paper, it is presented design and implementation of a system, based on the Polish language dictionary and the Agrovoc. Additionally some tests of implemented system are discussed.

Keywords: indexing, integrating sources of information, the Semantic Web, knowledge management.

1. Introduction

Nowadays, the ability to use data resources, information and available knowledge is crucial. Increasing technological capabilities makes informational resources to grow faster and faster. We have more and more information such as test results, descriptions of experiments or statistical data, which are difficult to

analyze without appropriate technological tools. However, the resources are more accessible than before, because resources are stored in digital form and a modern technology allows new methods of searching and analysis. Information systems have become indispensable in the process of acquiring, storing, processing and sharing of knowledge. This situation also applies to issues in the field of agriculture and life sciences. It is necessary to properly describe and classify resources, without the search even using modern tools is troublesome and time-consuming. The classification of scientific publications is easier, because they define the keywords; however the defined keywords are not always sufficient. The classification of Web pages or other resources poorly described is more difficult. In such case, the automatic indexation of information can facilitate the task. An automatic indexation allows defining relationships between documents and classifying them. It is used by popular search engines like the Google. However, the algorithms used in the search engines are based not only on the presence of words but also on the number of links between the web pages. The HTML has “meta” tag which can be used to specify page keywords and description of the document, for example: <meta name="keywords" content="potatoes, seed">. The metadata can be used by browsers, search engines, or other web services, but in practice they are ignored. One of the latest initiatives on the Internet is the microdata format, part of the HTML5 standard. This format uses the ontology which is available on the portal schema.org. Developers of search engines support this initiative focused on the most common search terms on the Internet like movies, concerts, etc. This ontology is prepared in the English language only, it does not include the concepts associated with agriculture. The closest to agriculture are recipes which are the frequent target of Internet search. The real objective of microdata is advertising; scientific papers and resources needed in science are rather not in the interests of advertisers.

It should be noted that description of the online resource at the semantic level is a matter under consideration for many years. An idea of Semantic Web has a long history. The methods of describing resources semantically were presented in the work [6].

The subject of our interest is the automatic indexing of resources according to selected set of concepts. The first goal was to examine how we can use existing tools to indexing agricultural texts. A lot of scientific papers in the field of agriculture are written in native languages. This also takes place in Poland. Indexing and more generally, knowledge extraction from documents is difficult in languages that have an extensive inflexion. Polish language is one of mentioned languages. The main objective of this study is preparing indexing tool for agricultural texts in the Polish language.

In the rest of this paper, firstly general methods for indexing texts are discussed. Next, the Agrotagger tool prepared by FAO is presented and its functionality is tested. Then based on the specificities of both the Polish language and the

field of agriculture design, implementation and functionality of a prototype system for indexing documents in Polish language are described.

2. Text indexation algorithms and tools

Indexing of documents is not a new issue in computing, it was often associated with problems of automatic text processing. Text processing is an old subject of computer science. It was issue of interest when the documents in electronic format were the only margin of information resources. Searching for information from text documents have been the subject of research in the field of natural language processing (NLP) and lately knowledge management (KM). We can specify that the main purpose of searching for information is finding the material (usually documents), which meets our information requirements of large collections usually stored on computers [9]. The indexation process is generally the first step of the searching for information in a given context and it is related to text representation. Because of that, the system can select and rang documents in accordance with the user requests. Our goal is close to named-entity recognition (NER) also known as entity identification or entity extraction. NER is a subtask of information extraction that seeks to locate and classify elements in text into pre-defined categories such as the names of persons, organizations, locations, etc. For us important is pre-defined set of concepts connected with agriculture and we want to rank documents according those concepts.

The most important techniques used by full-text indexing are: a part of the speech recognition (called part of speech tagging) and extracting the core of words (called stemming). Identifying parts of speech is described in [8], where it is concluded that currently in the English texts parts of speech recognition is quite accurate. Very useful is hidden Markov model part-of-speech tagging [5]. There were developed a lot of recognition algorithms for stemming, the best known are: Lovins algorithm [7], Porter algorithm [11], and Paice/Husk algorithm [10]; a comprehensive review of the literature can be found in the second chapter of the work [9]. Most of these methods work well in English but not in languages as Polish. Many attempts have been made to adapt these methods for Eastern European languages, e.g. [3]. There are a lot of publications about searching for scientific information and, in particular, the full-text indexing for scientific papers, such as [4], but they are generally devoted to specific issues. We have to note that there are commercial general purpose solutions such as Key Phrase Extractor business service Sematext, AlchemyAPI or Dandelion by Spaziodati. Academic projects mainly use non-commercial solutions such as http://labs.translated.net/terminology-extraction/ or http://texlexan.sourceforge.net/, but in general, they are good only for English language. There is interesting tool useful for text processing: Apache Lucene

(http://lucene.apache.org), part of it is Stempel - algorithmic stemmer for Polish language. Many specific tools for Polish language has been constructed as academic projects, an overview of these tools is available on the Computational Linguistics in Poland website. Unfortunately none of these tools is dedicated to the issues of agriculture.

3. Agrotagger

We are interesting in the text indexing of publications in the life sciences and especially in the agriculture. The FAO initiative Agrotagger [1] is very interesting because it uses a keyword extraction based on Agrovoc thesaurus [2]. FAO prepared first version of Agrotagger in collaboration with Indian Institute of Technology of Kanpur (IITK). Several versions were created, some based on keyword extraction algorithm engine and on reduced subset of Agrovoc called Agrotags (http://agropedialabs.iitk.ac.in:8080/agroTagger). Additionally MIMOS company in collaboration with IITK and FAO produced application on the top of the IITK tagging service by storing the generated keywords as RDF triples. Moreover, FAO has collaborated with the Metadata Research Center of the University of North Carolina who include Agrovoc along with a host of other thesauri in their indexing and browsing tool known as HIVE. Unfortunately, both last systems are periodically not available. Finally FAO assembled an Agrovoc-based indexing package using the Maui indexing framework (http://maui-indexer.appspot.com/). There is information on FAO web pages that source code can be accessed at GitHub but it is only available as console application under UNIX operating system.

A number of tests and experiments on a variety of documents in the English language were performed on two available versions of Agrotagger – Maui (Figure 1) and IITK (Figure 2). Four simple texts about potatoes were prepared as the basis for comparing mentioned above two services. The text 1 is about history of potatoes and generally about varieties of potatoes. The text 2 is generally about potatoes and their composition of the chemical elements and nutritional properties and about countries with biggest production of potatoes. The text 3 is a “Guidelines for Preventing and Managing Insecticide Resistance in Aphids on Potatoes”. The text 4 is about seed potatoes from Great Britain. The results of the study were a huge surprise despite the fact that the service uses the IIKT reduced set of concepts from Agrovoc. Most of the selected keywords were different. Concepts selected by both services are written in bold font in the table 1. It can be concluded that only IIKT service upheld some semantic relationships by adding the broader concepts (i.e. tracheophyta in Text1).

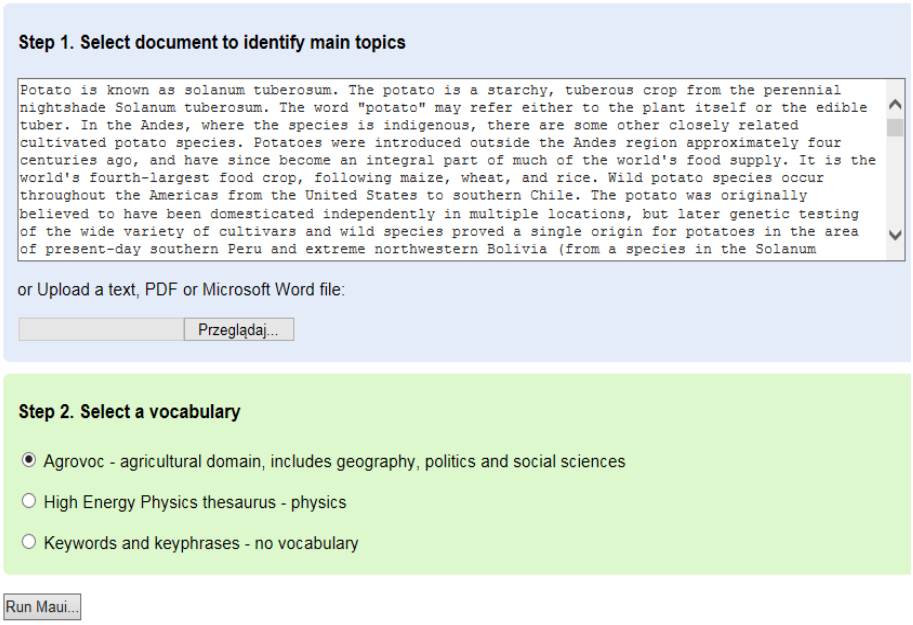

Figure 1. Maui indexing service

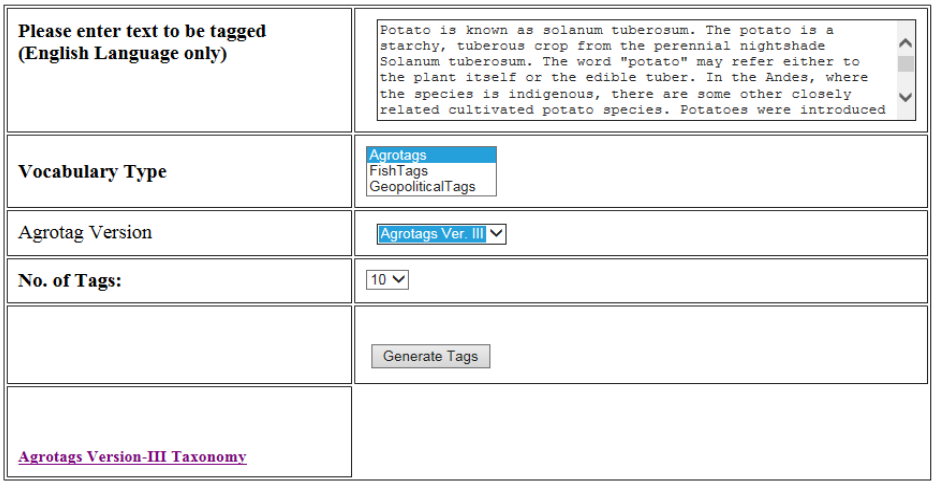

Figure 2. IITK indexing service

Unfortunately, Agrotagger analyzes only concepts from the English version of Agrovoc thesaurus, so for texts in the Polish language only the abstract in English is indexed. In conclusion we can say that, despite the fact that Agrovoc is a multi-lingual thesaurus, the indexation process is performed only in English and in its

current form is useful only for publication in English. In addition, the large differences in both the indexers show that it is necessary to analyze more precisely how indexing is performed.

Table 1. Comparing IITK and Maui indexing service

	IITK extracted keywords	Maui extracted keywords
Text 1	potatoes, organisms, processing, world, cooking methods, processed animal products, **varieties**, species, tracheophyta, brewing	Food crops, Vegetables, Food supply, Solanum tuberosum, Solanum, USA, Developing countries, **Varieties**, Perennials, Foods
Text 2	**potatoes**, world, processing, geographical regions, productivity, diseases, cooking methods, metallic elements, planting, crops	Livestock, **Potatoes**, Vegetables, High water, North America, Developing countries, Asia, Sweet potatoes, Diet, South America
Text 3	hexapoda, **potatoes**, **crops**, **insecticides**, mace, productivity, tracheophyta, **pests**, **species**, biopesticides	**Crops**, Horticulture, **Pests**, Risk analysis, **Species**, **Insecticides**, Aphidoidea, Control methods, Cereals, **Potatoes**
Text 4	plant production, **potatoes**, world, propagation materials, diseases, **varieties**, socioeconomic development, **crops**, planting, tracheophyta	Seed, **Crops**, Health, **Varieties**, Seed potatoes, Industry, Developing countries, Horticulture, Quality assurance, **Potatoes**

4. Prototype indexing system in Polish

Agrotagger is not relevant to the Polish texts. Because of that we decided to create our own system. The main requirement was formulated as follows: system have to index publications in Polish and eventually profile the results on the basis of the Agrovoc thesaurus. Ultimately, the system is expected to be similar to Agrotagger. Initially, the action was limited to the first requirement. Currently documents have to be in the txt format. To prepare a database of words with inflected forms open-source dictionary of Polish language (www.sjp.pl) was used. First prototype was designed in a client-server architecture. An additional requirement was the study of semantic relationships between publications. System, results and conclusions were published in the [12].

On the basis of the experience with mentioned system new version was designed (Figure 3). The main component of new version is Indexer application, Agrovoc thesaurus is accessed through the Web Service and the Polish Language Dictionary is used as local copy. In the current version only the files in text format are analyzed. Process of indexing is the following. Firstly the Polish Language Dictionary is loaded, analyzed and processed. After this process array with con-

cepts, numbers of inflection forms, grammatical categories and list of inflection form is prepared. During the second step document is loaded and after stemming process vector of words is constructed. In the third step nouns from vector of words are associated with the concepts from Agrovoc. At the end of the process the selected terms are saved in text file.

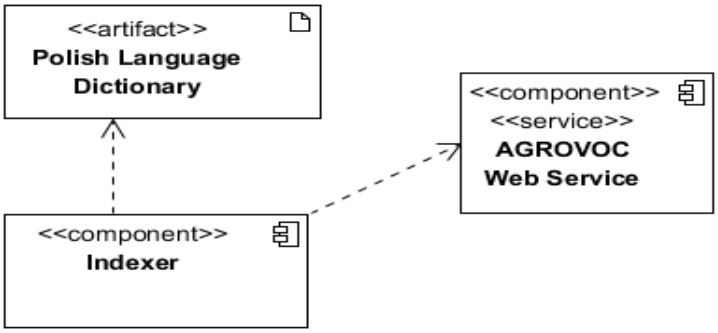

Figure 3. System architecture

More than a dozen Polish publications form Agricultural Engineering Journal (Inżynieria Rolnicza - IR) were indexed to test our solution. Selected publications were related to the cultivation and processing of maize. Six publications have been selected and results connected with them are presented in table 2. The basic information about selected publications is described below. “Text A” is “Information system for acquiring data on geometry of agricultural products exemplified by a corn kernel” (Jerzy Weres: „Informatyczny system pozyskiwania danych o geometrii produktów rolniczych na przykładzie ziarniaka kukurydzy”. IR 2010 Nr 7); “Text B” is “Assessment of the operation quality of the corn cobs and seeds processing line” (Jerzy Bieniek, Jolanta Zawada, Franciszek Molendowski, Piotr Komarnicki, Krzysztof Kwietniak: „Ocena jakości pracy linii technologicznejdo obróbki kolb i ziarna kukurydzy”. IR 2013 Nr 4); “Text C” is “Methodological aspects of measuring hardness of maize caryopsis” (Gabriel Czachor, Jerzy Bohdziewicz: „Metodologiczne aspekty pomiaru twardości ziarniaka kukurydzy”. IR 2013 Nr 4); “Text D” is “Evaluation of results of irrigation applied to grain maze” (Stanisław Dudek, Jacek Żarski: „Ocena efektów zastosowania nawadniania w uprawie kukurydzy na ziarno”. IR 2005 Nr 3); “Text E” is “Extra corn grain shredding and particle breaking up as a method used to improve quality of cut green forage” (Aleksander Lisowski, Krzysztof Kostyra: „Dodatkowe rozdrabnianie ziaren i rozrywanie cząstek kukurydzy sposobem na poprawienie jakości pociętej zielonki”. IR 2008 Nr 9); and “Text F” is “ Comparative assessment of sugar corn grain acquisition for food purposes using cut off and threshing methods”

(Mariusz Szymanek: „Ocena porównawcza pozyskiwania ziarna kukurydzy cukrowej na cele spożywcze metodą odcinania i omłotu". IR 2009 Nr 8).

Table 2. Comparing keywords, extracted keywords and Agrovoc keywords

	keywords	**extracted keywords**	**extracted Agrovoc keywords**
Text A	modelowanie geometrii, wykrywanie krawędzi, siatka strukturalna MES	Produkt, siatka, ziarniak, geometria, węzeł, obraz, system, współrzędna, element	Produkt, ziarniak, kukurydza, model, inżynieria
Text B	linia technologiczna, obróbka kolb kukurydzy, ziarno, jakość pracy	Ziarno, kolba, kukurydza, odmiana, jakość, praca	Ziarno, kukurydza, odmiana, jakość, praca
Text C	twardość, okrywa, zarodek, ziarniak, kukurydza	Twardość, wartość, pomiar, faza, ziarniak, czas, tkanka	Twardość, pomiar, ziarniak, czas, metoda, głębokość, wielkość, kukurydza
Text D	nawadnianie, kukurydza na ziarno, nawożenie azotowe, odmiana	Kukurydza, nawadniać, ziarno, odmiana, plon	Kukurydza, ziarno, odmiana
Text E	kukurydza, rozdrabnianie, toporowy zespół tnący, długość sieczki	Ziarno, kukurydza, rozdrobnić, wskaźnik, wartość, sieczka, długość, łopatka	Ziarno, kukurydza, długość, łopatka, roślina, sieczkarnia
Text F	kukurydza cukrowa, odcinanie, mrożenie, omłot, jakość	Ziarno, kukurydza, kolba, omłot, jakość, odmiana, odcinać	Ziarno, kukurydza, jakość, odmiana, metoda, masa

The first conclusion is that the analysis of the concepts (nouns) is not sufficient, it is necessary to take into account the verbs and adjectives and more specifically phrases. The results are generally interesting. In publication A author defined as keywords only terms connected with finite-element method. It is interesting that concept maize is only in title but not in the keywords, despite the publication refers to maize. Implemented Indexer relatively good recognized topics related to agriculture. In publications B, D and F situation is similar, but authors, in contrast to publication A, inserted not only the technological keywords. Indexer did not recognize vocabulary associated with technology but fairly well identified farming concepts. The best results were obtained for publications C and E. Additionally, the Agrovoc thesaurus lets us print all the broader concepts for example for kukurydza (maize) we have: Agrostidaceae, Andropogonaceae, Arundinaceae, Arundinellaceae, Avenaceae, Bambusaceae, Chloridaceae, Eragrosteae, Eragrostidaceae, Festucaceae, Trawy, Hordeaceae, Lepturaceae, Maydaceae, Melinideae, Oryzaceae, Panicaceae,

Phalaridaceae, Gramineae, Poeae, Sporobolaceae, Stipaceae, Tripsaceae, Zizanieae, Plewowce, Wiechlinowate. Analogously we can obtain narrower concepts: Kukurydza zwyczajna or Koński ząb (roślina).

5. Conclusions and future work

Food and Agriculture Organization prepared, as a part of Agricultural Information Management Standards initiative, Agrotagger, tool for indexing documents in the field of agriculture. Agrotagger uses Agrovoc multilingual thesaurus but is designed only for the English language. In addition, the process of indexing with concepts from the Agrovoc thesaurus is not precisely specified, different versions gives different results. In this paper we presented an approach for Agrovoc based indexing for text documents in Polish. The first prototype tests of Indexer allow us to determine that the results are promising. Indexing system takes on the case of publications in text format. It means that now we have to preprocess files in different formats, for example pdf files. The next step should be to enable direct action on documents in doc and pdf format and, above all, on the web pages. Moreover it is necessary to prepare the body of texts intended for systematic testing and interfaces for reading various formats of publications. The next direction of further development should be taking into account during indexing semantic connections as broader, narrower and related concepts.

REFERENCES

[1] AgroTagger. http://aims.fao.org/agrotagger (access 19.11.2014).

[2] AGROVOC, http://aims.fao.org/standards/agrovoc/about/ (access 19.11.2014).

[3] Dolamic, L. Savoy, J. (2008) *Stemming Approaches for East European Languages*. Advances in Multilingual and Multimodal Information Retrieval, Vol. 5152, 37-44.

[4] Gupta S., C.D. Manning, (2011) *Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers*, In Proceedings of the International Joint Conference on Natural Language Processing. http://nlp.stanford.edu/pubs/gupta-manning-ijcnlp11.pdf (access 19.11.2014).

[5] Jurafsky, D., Martin J. H. (2009) *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd ed. Prentice-Hall.

[6] Karwowski W., (2010) *Ontologies and Agricultural Information Management Standards.* Information systems in managment VI, ed. P. Jałowiecki & A. Orłowski, WULS Press, Warszawa 2010.

[7] Lovins, J. (1968) *Development of a Stemming Algorithm*. Mechanical Translation and Computational Linguistics 11(1-2), 11-31.

[8] Manning C.D., (2011) *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics*? Computational Linguistics and Intelligent Text Processing, 12th International Conference, Proceedings, Part I. Springer LNCS vol. 6608, 171-189.

[9] Manning C.D., Raghavan P., Schuetze H. (2008) *Introduction to Information Retrieval*, Cambridge University Press.

[10] Paice C., Husk G., (1990) *Another Stemmer*, ACM SIGIR Forum 24(3). 56-61.

[11] Porter, M. (1980) *An algorithm for suffix stripping*. Program 14(3), 130-137.

[12] Wrzeciono P., Karwowski W. (2013) *Automatic Indexing and Creating Semantic Networks for Agricultural Science Papers in the Polish Language*, Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual, Kyoto.