INFORMATION SYSTEMS IN MANAGEMENT

Information Systems in Management (2017) Vol. 6 (3) 192−202

METHODS OF AUTOMATIC TOPIC MINING IN PUBLICATIONS IN AGRICULTURE DOMAIN

WALDEMAR KARWOWSKI, PIOTR WRZECIONO

Department of Informatics, Warsaw University of Life Sciences (SGGW)

Today the vast majority of resources are available in digital form. Publications frequently are related to topics not set out in the title or even summary. In this paper we presented and discussed examples of methods of finding the common topic of a publication in the field of agriculture with the use of AGROVOC dictionary. The focus is on publications in the Polish language, and the possibilities of the use of the semantics defined in the multi-language thesaurus AGROVOC. First indexing tools, especially Agrotagger, which is useful for documents in the field of agriculture, are presented, and also the test results of Agrotagger are discussed. Next the semantic technologies implemented in the AGROVOC thesaurus are discussed. In the final part, we described the design and implementation of a system, based on Polish language dictionary and AGROVOC. Additionally some tests of implemented system are discussed.

Keywords: indexing, thesaurus, the Semantic Web, knowledge management

1. Introduction

Information technology has enabled both acquiring and storing data more than ever in the past. Nowadays, technological capabilities are increasing and make informational resources to grow faster every year. Direct analysis of such a large amount of information is not possible by traditional methods like reading. Fortunately today informational resources are in digital form and it is possible to search and analyze them with modern software tools. Analysis of numeric and text re-

sources is not an easy task, because they don't have a specific structure. Of course, much of the data is now stored in relational databases in which both the structure and query language are well defined. However, the number of poorly structured resources is definitely greater than the structured. Currently, most resources are Web pages available in HTML format. Therefore indexing of such resources is very important and can in a significant way speed up the searching. Some publications like scientific publications can be indexed easier, because they define the keywords, but keywords not common for other types of publications. For Internet pages we have, among others, the microdata format, part of HTML5 standard that allows to store basic information about the publication. Developers of search engines, like Google, support this initiative focused on the most common search terms on the Internet like: movies, concerts, etc. Automatic indexation is used by popular search engines, but results of the search engines, in general, are not enough for many purposes. It is necessary to construct indexers dedicated for a particular field, for example agriculture and life sciences.

Process of describing informational resources needs the correct set of keywords connected with the particular field. The source of such carefully selected sets of words and phrases of words can be controlled vocabularies, which are used to tag units of information. There are many forms of such vocabularies. One of them is taxonomy. Taxonomy is defined as a tree structure of concepts. More advanced is thesaurus. Thesaurus, a set of semantically and hierarchically related terms, traditionally understood as a collection of synonyms. Most advanced is ontology a formal representation of a certain branches of knowledge, which consists of the record collections of concepts and relationships between them, which can be used as a basis for inference about the properties of these ontology concepts [4]. Mentioned above microdata format uses simple ontology which is available on the portal schema.org. Unfortunately this ontology, prepared in English only, does not include the concepts associated with agriculture. Listed types of controlled vocabularies allow indexing at different levels of semantic. It should be noted that idea of Semantic Web is connected with description of the online resource at the semantic level. Many techniques of making semantically resource descriptions were presented in the work [4].

The subject of our interest is automatic indexing of text resources in Polish language in the domain of agriculture at the semantic level, based on the AGROVOC thesaurus. The aim of the paper is to analyze examples of methods of finding the common topic of a publication in the field of agriculture, and to present indexing tool for text documents in Polish in agriculture domain based on AGROVOC dictionary. In the rest of the paper firstly selected indexing tools, among them Agrotagger tool prepared by FAO, are presented. Then techniques for semantic description used in AGROVOC are discussed. In the next part some in-

dexing issues connected with languages that have an extensive inflexion, like Polish language, are discussed. Finally prototype system for indexing documents in agriculture domain in Polish language is described and tests of this system are presented.

2. Text indexation tools

Automatic text processing is one of oldest subject of computer science, and especially automatic indexing of documents is one of the areas of it. Growing data resources, particularly on the Internet, resulted that need for automatic indexing has grown. In particular, indexing in a given context is very important in the field of knowledge management. In knowledge management big role plays predefined set of concepts connected with particular domain, like agriculture. Thanks to it, the indexing system can select and rang documents in accordance with the user requests. We have to note that there are many 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. Unfortunately none of these tools is dedicated to the issues connected with agriculture. There are systems which support multiple languages; good example is Thomson Reuters Open Calais service. This service supports English, French and Spanish and, according to owner, offers the easiest and most accurate way to tag the people, places, companies, facts, and events. Service is free for small text documents with limited features. It is very effective system for financial deals especially for listed companies on Stock Exchange, but issues related to agriculture are absent. Open Calais uses Calais Semantic Tagging Ontology for the OneCalais service. This ontology is published and available while the specific algorithms are the company secrets. Although this service is a good example of a semantic service based on ontology.

In the field of agriculture and natural sciences, the most interesting indexer is Agrotagger [1]. Agrotagger is FAO initiative, which for keyword extraction uses AGROVOC thesaurus [2]. Agrotagger was implemented in few pilot versions described in [5]. Two of them were most important and available as online services: first based on Keyword Extraction Engine a reduced subset of AGROVOC (http://agropedialabs.iitk.ac.in:8080/agroTagger), and second using the Maui indexing framework (http://maui-indexer.appspot.com/). The use was free, but implementation details, especially algorithms, have not been explained. Unfortunately, now both systems are not available and it is not known whether they will be

started again. Currently, the only available version of the Agrotagger is a computer program written in Java available as command line application, which code can be accessed at GitHub. This application is based on the Maui, which uses software Weka which allows machine learning. Now the only available version of Agrotagger is based on 780 publications, tagged by specialists with AGROVOC concepts, on which the indexer has been trained. As a result indexer basically indexes after some subset of terms from AGROVOC connected with mentioned 780 publications. To compare the command line Agrotagger with previously available services, there was performed test on the same texts as described in [5]. Text 1 is about history of potatoes and generally about varieties of potatoes. Text 2 is generally about potatoes their composition of the chemical elements and nutritional properties and about countries with biggest production of potatoes. Text 3 is a "Guidelines for Preventing and Managing Insecticide Resistance in Aphids on Potatoes". Text 4 is about seed potatoes from Great Britain. The results of the study, compared with previous tests, are presented in the table 1.

Table 1. Comparing IITK, Maui and Agrotagger 780 indexing

	IITK	Maui with AGROVOC	Agrotagger base 780
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	Cooking, Developing countries, species (taxa), Species, **Varieties**, Potatoes, Productivity, products, Production, Crops
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	Cooking, Developing countries, **Potatoes**, Productivity, products, Production, Crops, Crop (bird), species (taxa), Species
Text 3	hexapoda, **potatoes**, crops, insecticides, mace, productivity, tracheophyta, pests, **species**, biopesticides	Crops, Horticulture, Pests, Risk analysis, **Species**, Insecticides, Aphidoidea, Control methods, Cereals, **Potatoes**	Seed, Seeds, Sowing seeding (sugar), United Kingdom, species (taxa), **Species**, **Potatoes**, Productivity, products
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**	Seed, Seeds, seeding (sugar), Sowing, **Varieties**, **Potatoes** Products, **Crops**, Crop (bird), United Kingdom

Common concepts selected by two services and command line application, are written in bold font in the table 1. Conclusion is the following: most of the selected keywords are different. From the results we can see that generally no semantic possibilities of the AGROVOC were used. Only IIKT service used one semantic relationship by adding the broader concepts (i.e. tracheophyta).

3. Semantic technologies in AGROVOC

One of our goals is to utilize semantic relations in AGROVOC to increase the accuracy of finding keywords. In the current section, we will present technologies that allow the use of semantic. AGROVOC is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment etc. It is published by FAO and edited by a community of experts. AGROVOC consists of over 32,000 concepts available in 27 languages, among them is Polish [2]. AGROVOC is used by researchers, librarians and information managers for indexing, retrieving and organizing data in agricultural information systems and Web pages. AGROVOC is organized as thesaurus. AGROVOC previously was available in many formats: as a text file or even relational database, currently it is available as an SKOS concept scheme.

We will now briefly describe the underlying AGROVOC technologies. As we mentioned earlier thesauri organize concepts as synonyms, broader and narrower terms identifying the relationships and dependencies between them. To ensure accuracy appropriate national and international standards were defined. Older thesauri standards like ISO 2788 and NISO Z39.19 - 2005 apply to terms rather than concepts. They use the equivalence relations USE and UF (used for), additionally three types of link are used to semantically relate terms: BT (broader term) and NT (narrower term) express that a term's meaning is more general than another's. RT (related term) is used when a non-hierarchical associative link holds between meanings, which can be useful for applications which exploit the thesaurus [3]. ISO 25964 is the newest international standard for thesauri, in which is more attention on the concepts not only on terms. Concepts are indexing in a given context, and label each of them with a preferred term. The main types of relationship between the concepts and between the terms are: equivalence (between synonyms and near-synonyms), hierarchical (between broader and narrower concepts) and associative (between concepts that are closely related in some non-hierarchical way). Nowadays, in the Internet era, the biggest popularity gain standards, based on XML, developed by W3C. SKOS (Simple Knowledge Organization System) is

an XML standard and more precisely is RDF/XML application. RDF (Resource Description Framework) is described as “language for defining structured web based ontologies which will provide richer integration and interoperability of data among descriptive communities”. RDF document stores triples: subject–predicate–object and has the form of a directed graph. We have to note that there are many possible RDF notations, but currently RDF/XML is the basic. SKOS, as we mentioned, is built upon RDF and is representation intended to define thesauri and controlled vocabularies using the RDF/XML notation analogously to ISO 2788, NISO Z39.19 and ISO 25964. The SKOS metamodel is broadly compatible with the data model of ISO 25964 and intends to allow easy migration of thesauri defined by standards such as ISO 25964. SKOS provides three properties to attach labels to conceptual resources: prefLabel, altLabel and hiddenLabel. Semantic relations in SKOS play an important role for defining concepts. There are three standard properties for relations: broader, narrower and related. SKOS provides several properties that map concepts between different concept schemes there are: exactMatch, closeMatch broadMatch, narrowMatch and relatedMatch. Moreover SKOS makes it possible to define meaningful groupings or “collections” of concepts. Additionally SKOS has specific properties, broaderTransitive and narrowerTransitive, because not always broader relation is transitive. It is easy to create additional properties and classes and attach them to the standard SKOS vocabulary elements by using the subPropertyOf and subClassOf properties from the RDF Schema vocabulary, because SKOS is based on RDF. AGROVOC, thanks to SKOS, allows to describe quite rich semantic, but that's not all. AGROVOC uses additionally agrontology specific vocabulary of relations. Agrontology relations (properties) are grouped into very powerful hierarchies. Property hasRelatedConcept has subproperties CausativeRelationships, QuantitativeRelationship, TaxonomicRelationship. For example CausativeRelationship has subsequence subproperties: actsUpon, affects, benefitsFrom and so on. There are properties intended to label relations for example such property hasSynonym has subproperties: hasBroaderSynonym, hasLocalName, hasNarrowerSynonym, hasNearSynonym, hasOldName, hasScientificName, hasTradeName, hasCommercialName, isLocalNameOf, isTradeNameOf, isOldNameOf, isDeprecatedNameFor, scientificNameOf (Figure 1). As we can see the semantics contained in AGROVOC is rich and the possibilities of subtle differentiation are large.

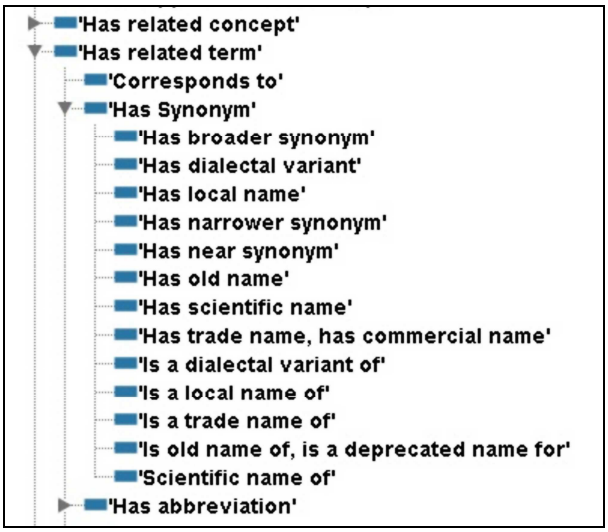

Figure 1. Agrontology label hasSynonym
Source: own preparation with Protégé editor

4. Indexing system in Polish

Although the AGROVOC is a multilingual thesaurus, Agrotagger analyzes only concepts from the English version. Because the indexation process is performed only in English, current form of Agrotagger is useful only for publication in the English language. The prototype indexing system relevant to texts in Polish was created and its functionality was described in [5]. Indexing system in Polish is based on database of words with inflected forms from open-source dictionary of Polish language (www.sjp.pl) and full version of AGROVOC. Prototype was designed in client-server architecture, AGROVOC thesaurus is accessed through Web Service, Polish Language Dictionary is used as local copy. System is a continuation of the application described in [6]. To take advantage of the semantic abilities of AGROVOC in addition to the stemming also was added analysis of semantic relations. The algorithm is as follows:

- As a first step we perform searching for the words (including the inflectional analysis of words), candidates for keywords, and we verify as to the presence of candidates in the AGROVOC thesaurus.
- In the second step we are looking for "related" concepts in the analyzed text. Currently "related" means such properties as related, prefLabel, altLabel, broader and isPartOf label.
- In the third step, the results of "related" concepts are added to the results obtained in the first step.

To compare the results with the previous version of the indexer the same six publications tested in [5] have been selected and indexed. The selected publications in Polish are from Agricultural Engineering Journal (Inżynieria Rolnicza - IR), and are related to the cultivation and processing of maize. “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).

The results of the test is presented in Table 2 – Table 7. The measure is the frequency of word for simple extractions and word and “related” concepts for extraction with semantic support.

Table 2. Text A Comparing keywords, extracted keywords and Agrovoc keywords

	extracted Agrovoc keywords	extracted Agrovoc keywords with semantic support
produkt	13%	13%
ziarniak	9%	10%
kukurydza	5%	5%
model	4%	4%
inżynieria	4%	4%
metoda	3%	6%

Table 3. Text B Comparing keywords, extracted keywords and Agrovoc keywords

	extracted Agrovoc keywords	extracted Agrovoc keywords with semantic support
ziarno	18%	23%
kukurydza	10%	10%
odmiana	9%	10%
jakość	6%	8%
praca	6%	8%
wilgotność	4%	6%

Table 4. Text C Comparing keywords, extracted keywords and Agrovoc keywords

	extracted Agrovoc keywords	extracted Agrovoc keywords with semantic support
twardość	20%	22%
pomiar	13%	14%
ziarniak	11%	11%
czas	11%	11%
metoda	3%	5%
głębokość	3%	4%

Table 5. Text D Comparing keywords, extracted keywords and Agrovoc keywords

	extracted Agrovoc keywords	extracted Agrovoc keywords with semantic support
kukurydza	16%	16%
ziarno	13%	15%
odmiana	7%	8%
Polska	4%	4%
roślina	3%	3%
temperatura	2%	2%

Table 6. Text E Comparing keywords, extracted keywords and Agrovoc keywords

	extracted Agrovoc keywords	extracted Agrovoc keywords with semantic support
ziarno	18%	18%
kukurydza	16%	16%
długość	10%	10%
łopatka	10%	10%
roślina	9%	9%
sieczkarnia	8%	9%

Table 7. Text F Comparing keywords, extracted keywords and Agrovoc keywords

	extracted Agrovoc keywords	extracted Agrovoc keywords with semantic support
ziarno	36%	37%
kukurydza	14%	14%
jakość	9%	11%
odmiana	8%	8%
metoda	6%	9%
masa	5%	5%

We have, not big but significant, improvement of results. Additionally we can see that it is necessary to take into account not only nouns but the verbs and adjectives, more specifically phrases. Currently AGROVOC contains some phrases, right now in our system phrases are divided into separate words.

5. Conclusions and future work

Agrotagger, tool for indexing documents in the field of agriculture, was implemented by Food and Agriculture Organization, as a part of Agricultural Information Management Standards initiative. Agrotagger is designed only for the English language, although the AGROVOC is a multilingual thesaurus. In addition, the current version is limited to a subset of the keywords from a set of 780 documents. Agrotagger does not make use of semantic relationships available in the AGROVOC thesaurus. In this paper we presented indexing tool for text documents in Polish in agriculture domain based on AGROVOC. Agrotagger in the current version does not take into account the rich semantic relations contained in the AGROVOC, the indexing system in Polish uses selected semantic relations from the thesaurus and allows more precise classification. The tests of indexing system show that the results are promising. Currently indexing system takes on the case of publications in text format, the next step should be to enable direct action on documents in doc and pdf format and, above all, on the web pages.

REFERENCES

[1] AgroTagger. http://aims.fao.org/agrotagger (access 19.11.2016).

[2] AGROVOC, http://aims.fao.org/standards/agrovoc/about/ (access 19.11.2016).

[3] Dextre Clarke S. G., Lei Zeng M., (2012) *Standard Spotlight: From ISO 2788 to ISO 25964: the evolution of thesaurus standards towards interoperability and data modeling*. Information Standards Quarterly Winter 2012, v. 24, no. 1.

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

[5] Karwowski W., Wrzeciono P., (2014) *Automatic indexer for Polish agricultural texts*. Information Systems in Management 2014, Vol. 3, no. 4, pp. 229-238.

[6] 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.