INFORMATION SYSTEMS IN MANAGEMENT

Information Systems in Management (2014) Vol. 3 (3) 159−169

THE APPLICATION OF ONTOLOGY FOR INDEXING OF PUBLICATIONS IN THE LIFE SCIENCES

WALDEMAR KARWOWSKI

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

Ontologies recently have important role especially in knowledge management systems dedicated for agriculture. In the paper, issues related to indexing documents against the ontology, are presented and discussed. Problems with indexing documents in Polish language which has an extensive inflection are described. There are presented and discussed examples of ontologies and thesauri in the field of life sciences, in particular possible to use to describe aspects of plant production. We have tested Agrotagger the existing tool for indexing agricultural texts with publication in Polish. Original software developed for indexing web pages in Polish against potato ontology is described. In the final part some conclusions and plans for further research are formulated.

Keywords: knowledge management, ontologies, text indexing, agriculture

1. Introduction

Nowadays access to information becomes of great significance. At the same time data production is growing much faster than ever before. It is necessary to find useful information from the growing data resources. An increasing role of many employees is collecting, organizing and utilizing information. Nowadays the vast majority of the information is in digital form and computers are used for the processing of information, but finally a man has to draw conclusions and make decisions. However, it is possible to use appropriate decision support software to

support the undertaking of a decision. Spreadsheet software is an example of commonly used application in many areas.

The subject of our interest, first and foremost, is searching of information from the Internet pages in order to indexation. Many scientific publications are available online, some of them, primarily research articles are usually described by keywords. Of course keywords make easier to find information, but they are not always sufficient. Moreover scientific papers are not currently the only source of information even in science. There is a growing amount of scientific information, such as measurement data, experimental data or statistical data described rather with their associated metadata than keywords. Many publications are a white papers, technical reports or descriptions of technology. Furthermore, the researchers, for sharing knowledge, use modern internet platforms like content management systems, blogs and social networking to present and discuss results before they are published in scholarly journals or after they have been published.

Our goal is to extract words from the content of the publication which may indicate that the publication concerns issues of our interest. To achieve the objective, it is necessary to define the vocabulary that describes the field of interest, in our case crop production and more general agriculture. Such methods and formats must take into account the context and semantics. On the other hand, it is necessary to prepare tools to analyze text in natural language taking into account the flexion. Since we are indexing texts in Polish it means that both a description and inflection analysis must be in Polish.

In the following chapter of this paper, we shortly present similar works. Next methods of text analysis taking into account the inflection will be presented. Then we introduce the methods useful for description of domain – agriculture and crop production; we will focus primarily on thesauri and ontologies. The fifth chapter presents existing solutions for indexing publications in the field of agriculture in English. Next chapter presents the concept of a prototype system for indexing publications in Polish language regarding the potato ontology. At the end conclusions and plans for the future will be presented.

2. Related works

The use of ontologies in the text analysis and processing has a fairly long history. One of the issues is the assessment of the similarity between documents and text document clustering. Hotho et al. [6] started from VSM (Vector space Model) for document and made concept selection and aggregation. On this basis ontology was constructed, and then they modified VSM measure between documents according to ontology. It means that ontology was constructed from analyzed documents. Comparing among vector representation approach, latent semantic indexing method and ontology based method was performed in [16]. Ontology based method

is that new resources registered within the system are linked to concepts from this ontology. In such a way resources may be retrieved based on the associations and not only based on partial or exact term matching as the use of vector model presumes. Authors concluded that the results were promising. In [7] it was studied similar clustering according to WordNet lexical database. Authors designed a new data model (considering the correlation between terms) on which the Euclidean distance measure can be used. Additionally modified measures taking account related concepts from WordNet with the weight 0.8 were used. In [8] a system for ontology based annotation and indexing of biomedical data is presented. The key functionality of this system is to provide a service that enables users to locate biomedical data resources related to particular ontology concepts. The system is integrated with NCBO BioPortal (http://bioportal.bioontology.org/annotator) and its objective is to annotate a large number of biomedical resources and to provide an index up to date of annotated resources elements. The system is based on a domain knowledge representation schema in form of ontology. The user can select multiple ontologies in different formats (OBO, OWL, etc.) in mentioned field. Authors noted that the system selects the appropriate terms in the given ontologies but detailed indexing algorithm has not been presented. Approach for indexing web pages using HTML tags are presented in [5]. The document is segmented due to the HTML tags <title>, <h1>, <h2>; weights are assigned depending on the importance of tags. Indexation is performed against concepts from Agrovoc thesaurus independently to every segment. The experiments demonstrated that the proposed approach was capable of automatically annotating segments with concepts that describe a segment's content with a high degree of accuracy. This publication is interesting because indexed text is in Arabic language. At the end of this short review we can conclude that we have a lot of work according to the English language, they are based on the traditional methods modified by the use of ontology. The works relating to other languages than English are quite rare.

3. Text indexing method and tools

Searching for information from text documents, have been the subject of research in the field of natural language processing (NLP) and, more recently, knowledge management (KM). We can specify that the main purpose of information retrieval system is finding material (usually documents) that meets our requirements information from large collections (usually stored on computers) [14]. Searching for information is depended on the document representation (flat files in many formats like text, pdf, Word doc; semi structured files like HTML, XML or documents in more structured forms in databases etc.) and the method of access to it. Text indexing is part of the process of information retrieval in a given context. Indexation process is generally the first step of the process; thereby the search-

ing/indexing system can select and rang documents according to the user's query. The main techniques used for indexing is a part of speech recognition and the core of word identification called stemming. There are many algorithms created to recognize the core of word, the best known are: Lovins algorithm [12] Paice/Husk algorithm [15] and Porter algorithm [17]; an extensive review of the literature can be found in the second chapter of the book [14]. We have to note that most of these methods work well in English, but not in languages with complex inflection like Polish. There have been many attempts to adapt mentioned methods to Eastern European languages for example [2] however, the results are not satisfactory. The part of speech recognition is another important technique and it is described for example in [13]. Today part of speech recognition for English texts is quite accurate. There are many other works of scientific information retrieval and indexing, devoted to specific issues [4].

In order to make indexing, it is possible to use existing commercial solutions such as Key Phrase Extractor by Sematext or service offered by AlchemyAPI. In the academic projects there are mainly used non-commercial solutions such as http://labs.translated.net/terminology-extraction/ or http://texlexan.sourceforge.net/. Such free available solutions are mainly prepared for English or very specific languages such as Catalan http: //www.uoc.edu/serveilinguistic/home/index.html. It is possible to develop own algorithms specialized for a particular purpose, and as is often done for a variety of issues.

4. Ontologies and thesauri in life sciences

Indexation can rely on selecting the most frequent words but generally it is not sufficient. It is necessary to define set of words related to the topic. To describe the fragment or whole domain we can use ontologies. The subject of ontology is the study of the categories of things that exist or may exist in some domain. Sowa [18] notes that without ontology, the terms and symbols are ill-defined, confused, and confusing. Definition of ontology used in computer science and in knowledge management, was formulated by Gruber: "An ontology is a formal specification of a shared conceptualization" [3]. This definition is very general and many types of domain description are included in it. A formal ontology should be specified by a collection of names for concept and relation types organized in a partial ordering by the type-subtype relation. An informal ontology may be specified by a set of types that are defined only by statements in a natural language. Taxonomies, controlled vocabularies and thesauri are examples of tools for less formal ontologies. They have been used for years in life sciences, librarianship or linguistics. It was necessary popularity to define appropriate standards for creating ontologies; nowadays most popular are the standards based on XML syntax defined by WWW Consortium - RDF (Resource Description Framework), OWL (Web Ontology Lan-

guage) which is an extension of RDF and SKOS (Simple Knowledge Organization System) designed for representation of thesauri which is based on RDF. More about thesauri and ontologies standards is presented in [9,10].

Ontologies are widely used in the life sciences, the most important examples of applications are given in the papers [9,10,11]. Ontologies help us to organize the knowledge contained in the publications and they are an essential component of knowledge management systems. In the domain of our interest very important are Plant Ontology (http://www.plantontology.org/) and Crop Ontology (http://www.cropontology.org/). The main goal of the Plant Ontology project is to provide controlled vocabularies for the plant-specific knowledge domains: plant anatomical entities and plant structure developmental stages. Anatomical controlled vocabulary describes plant's morphological and anatomical structures representing organ, tissue and cell types and their relationships. The second controlled vocabulary describes growth and developmental stages in various plants and their relationships, examples are germination, seedling, flowering, etc. The Crop Ontology current objective is to compile validated concepts along with their interrelationships on anatomy, structure and phenotype of crops, on trait measurement and methods as well as on germplasm with the multi-crop passport terms. Unfortunately both plant and crop ontologies do not have terms in Polish. Plant Ontology has only Japanese and Spanish version, Crop Ontology is only in English. For us this means that we can only follow those ontologies and, if necessary, create Polish versions. For our purpose more interesting is thesaurus Agrovoc - a multilingual glossary in SKOS format in the fields of agriculture, forestry, fisheries, food and other related fields developed by FAO [19], because it is available also in Polish.

5. Agrotagger and Annotator

In the area of agriculture most interesting initiative is Agrotagger developed by FAO [1]. It is a keyword extractor that uses the Agrovoc thesaurus as its set of allowable keywords. Agrotagger began as a collaboration with Indian Institute of Technology of Kanpur (IITK). Building on top of the popular Keyword Extraction Engine (KEA) the team created several versions, some based on a reduced subset of Agrovoc and others using the full set of Agrovoc concepts. MIMOS in collaboration with IITK and FAO produced an interesting application on top of the IITK tagging service by storing the generated keywords as RDF triples and building from this a tag cloud showing the most commonly extracted keywords. In addition, 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.

We tested all mentioned versions of Agrotagger. For the test, we used an article in Polish with English summary "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". Inżynieria Rolnicza. 2010 Nr 7*). In practice, Agrotagger (IITK) has taken into account only English words those encountered in the abstract and bibliography. There are: Image processing, Kernels, Triticum aestivum, Engines, Wheats, Models, Wood, Fruit, Processing, Drying. Similarly, Agrotagger in the version of MIMOS produced the same output (Fig.1).

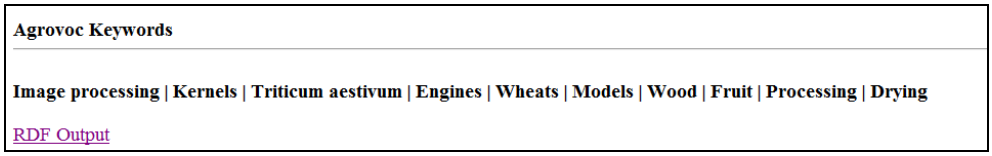

Figure 1. Result of Agrotagger indexing

Additionally this version made possibility to download result in RDF format (Fig.2). It should be noted that numbers in RDF output, for example mytermcode=25387, mean the indexes of concepts in the Agrovoc thesaurus (25387 is the code of Kernels concept).

Figure 2. Agrotagger indexing result as RDF file

Last tested tool, HIVE indexer, produced as result: Zea mays, Triticum aestivum with bigger font and Image processing, Kernels, Maize oil, Soft Wheat, Models, Maize, Wheats, Engineering (Fig.3).

Figure 3. HIVE indexing with Agrovoc result

The results are slightly different than in Agrotagger, which means that the HIVE indexer used semantic relationships (in this case synonyms) from Agrovoc during indexing process. At the same time, this means that although Agrovoc comprises semantics both Agrotagger versions did not use this. Moreover HIVE indexer presented results in the form of cloud tags, it means that more frequent concepts were written in bigger font.

We have to note that all three mentioned services in recent months were unavailable although links to them are available from AgroTagger page (all presented tests were made on April 2013). At present (November 2014) available is only version of Agrotagger (IITK) with reduced vocabulary named Agrotags (http://agropedialabs.iitk.ac.in:8080/agroTagger/index_PDF.jsp). Agrotags is the subset of Agrovoc. Agrovoc has about 40,000 agricultural concepts and Agrotags has only around 3057. The same publication gives the following tags: processing, data processing, plant products, plant oils, productivity, layering, agricultural products, drying, agricultural engineering, engineers. The result is different from the previous but the cause is the limited version of the thesaurus. Finally, within the context of the agINFRA project, FAO assembled an Agrovoc-based indexing package using the Maui indexing framework. There is information on FAO web pages that source code can be accessed at GitHub. Application is available to download as command line application under UNIX operating system.

An interesting tool for us is, mentioned earlier, BioPortal annotator (http://bioportal.bioontology.org/annotator) which uses, among other, Plant Ontology and Crop Ontology. Because the testing texts in the Polish language was meaningless we tested only English abstracts of papers from Agricultural Engineering (*Inżynieria Rolnicza*) against mentioned ontologies. The results were rather not interesting but allowed us to get an idea how Annotator uses ontologies.

In conclusion of this part we can say that, although Agrovoc is a multilingual thesaurus presented indexation process is conducted only in English and in its current form is not very useful for publication in Polish. The second conclusion is that searching algorithms are not documented, results for different versions differ. It means that constructing new algorithms is reasonable and testing them on wide variety of texts is necessary. Additionally it was found that indexing texts in Polish language requires ontologies prepared in Polish.

6. Prototype indexing system in Polish

Conclusions from the previous part of our paper justify the need for preparing indexing system for Polish language, because in the field of agriculture, multiple publications are in Polish. The main objective of prototype indexing system in the Polish language was to index web pages relative to the sample ontology. In our system indexing is made according to terms of potato ontology prepared in OWL. This prototype potato ontology is described in [9], it is important that ontology is designed in Polish language. Ontology is small and does not include synonyms and broader concepts. In the current version document is not segmented due to the HTML tags. Text is only filtered, which means that all tags and JavaScript codes are removed. To support inflection we used dictionary of Polish language http://www.sjp.pl/, which contains the inflected forms of Polish words. This dictionary of Polish language was useful but there are some specific terms that are not in it, like "rizoktonioza" (this concept has inflected forms in Polish and additionally may appear in Latin form: "rhizoktonioza"). In such situation we prepared set of ontology concepts inflected forms and attached this set of inflected forms for all concepts occurring in the ontology as fixed file. It is reasonable because ontology is not changed during indexing. As a consequence in our system filtered text is not transformed according to inflection. We have to note that indexing system for texts in Polish was presented in [20]. In mentioned system the concepts from text are dynamically transformed into basic forms during indexation process, according to Polish language dictionary. Our approach is enough because there are only a few important classes in potato ontology: potato (*ziemniak*), component (*składnik*), product (*produkt*), disease (*choroba*), pest (*szkodnik*), disease protection product (*środek chorobobójczy*) and insecticide (*środek owadobójczy*). In addition only potato class name is strictly connected with our issue, other class names are more general. For this reason, we have to search in the indexed text only instances (individuals) of those classes. It means that we do not search word component but rather words water (*woda*) or (*skrobia*) which are instances of component. As a basic measure of correspondence we took frequency of words in a text. Additionally weights were connected with every word: 1 for potato and its individuals; 0.5 for component, product, disease or pest individuals; 0.25 for disease protection product and insecticide individuals. An example results are presented in table 1.

The results show that popular sites about the potato have the highest compatibility with the potato concept. Web pages of research institutes and pages with professional knowledge about potato have less compatibility. Web page of deputy named Ziemniak has a relatively low compliance with our issue.

Table 1. An example results of page correspondence with potato

WWW page	Correspondence (in promiles)
http://groole.pl/o-ziemniakach	80,64516
http://www.polskiziemniak.pl/	64,03941
http://www.ihar.edu.pl/ziemniak.php	34,95935
http://pl.wiktionary.org/wiki/ziemniak	51,09489
http://agricopolska.pl/index.php/odmiany/jadalne	22,38254
http://www.ziemniak.pl/	9,02935

7. Conclusions and future work

FAO on the portal of the Agricultural Information Management Standards presents an AgroTagger, tool for indexing documents in the field of agriculture, which is designed for the English language. Tests have shown that in such form a tagger is unsuitable for indexing documents in Polish language. Agrotagger uses only the Agrovoc thesaurus. BioPortal Annotator indexes against many ontologies but also is useless for Polish texts. In this paper we presented an approach for ontology-based indexing for web pages in Polish. The first results of the prototype indexing application are interesting however, it is necessary to perform a more systematic study of web pages related to agriculture. On the basis of bigger set of examples it will be possible to improve the weights assigned to the concepts connected with main concept. Ultimately, it is necessary to prepare the corpus of texts in html format for systematic testing, which would allow further improvement of the system. The first extension can be segmentation page content due to the HTML tags <title>, <h1>, <h2> and assign weights depending on the importance of tags. It seems reasonable combination of our system with the web crawler to index the linked page groups. On the other hand, we have to prepare the interface for documents in format other than HTML, in the first place in doc/docx and PDF formats. Although it is used only ontology for one vegetable the proposed approach enables to adapt the system to new ontologies. In the future it is planned extension of the ontology with additional concepts. In parallel Agrovoc thesaurus will be used in order to complete the concepts with broader and narrower terms. After such improvements the application can be practically used for automatic indexing of texts.

REFERENCES

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

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

[3] Gruber, T., (1993) A translation approach to portable ontologies. Knowledge Acquisition, 5(2), 199-220.

[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] Hazman M., El-Beltagy S.R., Rafea A. (2012) An Ontology Based Approach for Automatically Annotating Document Segments. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 2, March 2012.

[6] Hotho, A., Maedche, A., & Staab, S. (2002) Ontology-based text document clustering. KÜNSTLICHE INTELLIGENZ 16(4), 48-54.

[7] Jing, L., Zhou, L., Ng, M. K., Huang, J. Z. (2006) Ontology-based distance measure for text clustering. In Proc. of SIAM SDM workshop on text mining, Bethesda, Maryland, USA.

[8] Jonquet C., Musen M.A., Shah N.H. (2008) System for Ontology-Based Annotation of Biomedical Data. International Workshop on Data Integration in the Life Sciences, DILS'08. 2008, Springer Lecture Notes in BioInformatics 5109, 144–152.

[9] Karwowski W. (2013), Design and implementation of ontology for plant production, Information systems in management XVIII / sci. ed. Piotr Jałowiecki, Arkadiusz Orłowski. - Warsaw: WULS Press 2013, 79-90.

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

[11] Karwowski, W. (2010) Standards based on XML in agricultural knowledge management systems. Informatyka ku Przyszłości, Warszawa. (in Polish)

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

[13] 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.

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

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

[16] Paralic J., Kostial I. (2003) Ontology-based Information Retrieval. In: Proc. of the 14th International Conference on Information and Intelligent systems, 23-28.

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

[18] Sowa John F. Semantic networks, http://www.jfsowa.com/pubs/semnet.htm (access 19.10.2013).

[19] Tezaurus Agrovoc. http://aims.fao.org/standards/agrovoc/about/ (access 19.11.2014).

[20] 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.