Scientific quarterly journal ISNN 1429-7264

Agricultural Engineering

2014: 4(152):103-110

Homepage: http://ir.ptir.org

DOI: http://dx.medra.org/10.14654/ir.2014.152.85

AUTOMATIC INDEXING OF INFORMATION RESOURCES CONCERNING AGRICULTURE IN POLISH

Waldemar Karwowski[*], Piotr Wrzeciono

Department of Computer Sciences, Warsaw University of Life Sciences in Warsaw

[*]*Contact details: ul. Nowoursynowska 159, 02-776 Warszawa, e-mail: waldemar_karwowski@sggw.pl*

ARTICLE INFO

Article history:
Received: March 2014
Received in the revised form:
September 2014
Accepted: October 2014

Keywords:
indexing
integrating sources of information
semantic network
knowledge management

ABSTRACT

Contemporary research and production activity require searching and collecting a variety of information, this also applies to issues in the field of agriculture. Today, the vast majority of resources are available in a digital form. 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. Extraction of knowledge is not very convenient in languages such as Polish language with a very extensive inflection. In Polish, the following parts of speech inflect: verbs, nouns, numerals, adjectives, and pronouns. Proper indexing requires an initial reduction of grammatical forms, to which the authors have used the dictionary of the Polish language and have developed a programme of reducing. Moreover the algorithms for determining weights corresponding to the validity of the appointments taking into account the prevalence of terms and their position in the document were developed and implemented.

Introduction

In the modern world information, knowledge and skill of using available resources of data is significant. Even greater technological possibilities cause that information resources are growing faster and are at the same time more available, on the other hand modern technology enables their searching and analysis. We have more and more such information as research results, description of experiments or sets of statistical data, which are difficult to be analysed without appropriate technological tools. It may be said that IT systems have become indispensable in the processes of searching, storing, processing or making knowledge available; such situation concerns also issues from the field of agriculture. Presently, publications of articles, research results or results of executed projects are prepared in an electronic form and Internet is widely applied for their availability. A paper form still frequently occurs parallel with a digital form, but it may be said that it has become secondary. Simultaneously, a reverse process is going on, paper publications prepared many years ago are digitized in order to facilitate their access. Description and proper classification of resources, without which searching even with the use of modern tools is difficult and time consuming, is indispensable.

Classification of scientific publications is easier, they define key words, they determine in bibliography references to other publications; however, such information is not always sufficient. Classification of internet websites or other resources which are weakly described is slightly more difficult. In such cases, automatic indexation of included information is useful. It also enables determination of relations between documents through automatic generation of semantic relations which is useful also for scientific publications. Automatic indexation is used by popular search engines - robots which are scanning a global network all the time are a foundation of success of Google company. However, classifications used in search engines are based on the number of connections between pages and rather mechanical indexing of words which occur on these pages. It should be noted that there have been attempts of semantic description of resources, in particular for web pages, these issues were presented for example in the paper written by Karwowski (2010). In order to describe resources, a standard of metadata Dublin Core, of however more general nature, was defined. Presently, one of the newest initiative concerning the Internet is a microdata format which is a composing part of HTML5, onthology prepared for this format is available on schema.org portal. Creators of search engines, who support the initiative focused on issues, which are searched for the most often in the Internet: films, concerts etc. This ontology prepared only in English, does not include concepts related to agriculture, the most close to agriculture are culinary recipes, which are often searched for by the Internet users. Concepts which enable placing advertisements prevail; scientific papers and resources necessary in learning are not of the most interest for advertisers.

Indexing documents is not a new issue in informatics; it has been frequently related to the problems of automatic translation of texts. It was dealt with when documents in electronic form constituted only a margin of informative resources. Presently this issue, in the times of the Internet gets again an important meaning, analysis of texts and searching for information is commonly used, for example in the scope of management of knowledge. Investigation of semantic relations between concepts gets more significance, which allows the use of automatic concluding methods. Indexing, a more generally extracting knowledge from documents is difficult in languages which have an extensive inflection. Polish language is one of them. The following parts of speech are inflective: verbs, nouns, numerals, adjectives and pronouns. Moreover, the number of forms is high. Some verbs have even over one hundred inflectional forms (Słownik Języka Polskiego) formed based on an infinitive. Due to extensive inflection, the process of automatic indexing prepared for English used for Polish generates many "artificial" concepts, which in reality are examples of one concept. Many scientific papers in agriculture are written in mother tongues. It mainly results from the fact that papers in this field often refer to many problems specific for each country. Such situation takes place also in Poland.

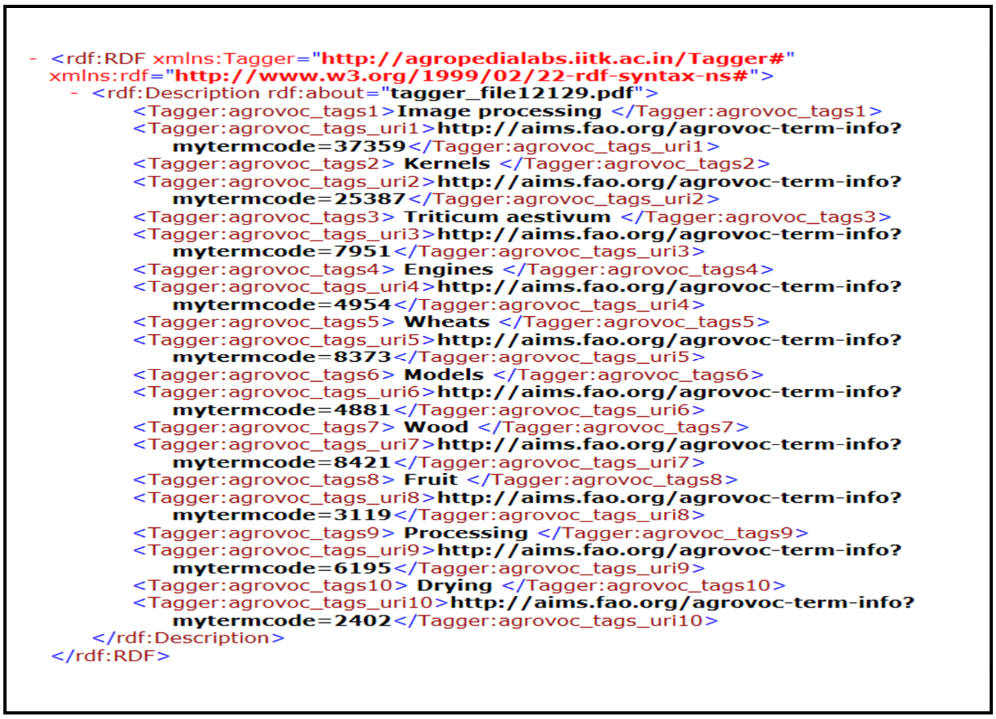
The initial objective of this paper was to investigate how to use existing tools to index texts related to agriculture in Polish. Then, based on specificity of both language as well as a field, a prototype of author's system for indexing documents, was designed and implemented. The paper presents results of contemporary research and conclusions developed based on the obtained indexing results.

Tools for indexation of texts

Issues of information retrieval from text documents, as mentioned above, have been the object of research in the field of processing a natural language and more up-to-date management of knowledge for many years. Information retrieval is related to its representation and manners of storing and access to it. It may be said that the main objective of the system of information retrieval is "finding material (usually documents), which meets our information requirements from among big collections (usually stored in computers)" (Manning et al., 2008). Indexing a text is a part of the process of information retrieval in a given context. The indexation process is generally the first step of this process, thanks to this process the system may select and rank documents according to the users' question. The most important techniques applied at indexing is *part of speech tagging* and recognition of stems occurring in inflection (English *stemming*). Many algorithms of stemming have been developed. The most popular are: algorithm by Lovins (Lovins, 1968), Paice/Husk (Paice and Husk, 1990) and Portera (Porter, 1980);an extensive review of literature may be found in the second chapter (Manning et al., 2008). Majority of methods can be easily used only in English, thus attempts of adjusting these methods to the Eastern Europe languages, have been made (Dolamic and Savoy, 2008). On the other hand, recognition of speech parts is described for example in Manning's paper (2011), on this basis it may be said that presently in English texts, parts of speech tagging is quite precise. Recently many papers devoted to the issue of scientific information retrieval and in particular to indexing scientific works, have been written, they are devoted generally to specific issues (Gupta and Manning, 2011). In order to carry out indexing, the use of existing commercial solutions such as Key Phrase Extractor by Sematext or service by AlchemyAPI, is possible. In academic designs, mainly non-commercial solutions or demonstrative solutions are used, such as Translated Lab Terminology Extraction (http://labs.translated.net/terminology-extraction/) or project TexLexAn (http://texlexan.sourceforge.net/). Similarly to previously mentioned paper, they concern English or specific languages, such as Catalan (http://www.uoc.edu/serveilinguistic /home/index.html). Developing own algorithms specialized for realization of a specific aim is also possible; research in this field were carried out for Polish (Branny, 2005).

This paper tackles the issue of indexing text in publications on agriculture and more generally in life sciences. In this field, AgroTagger, which for key words extraction uses (Thesaurus AGROVOC) as a set of admissible key words, is a very interesting initiative undertaken by FAO

As a part of this initiative, an internet system was developed in Indian Institute of Technology Kanpur (moreover, an analogous system is created in cooperation with MIMOS company; unfortunately both those systems are periodically unavailable). AgroTagger extracts the most important concepts and presents them in RDF format (fig. 1). Unfortunately, AgroTagger analyses concepts from English version of thesaurus AGROVAC (AGROVAC is a multi-language dictionary, includes concepts also in Polish), thus, an abstract of an article written in English and English words placed in the text are decisive. Figure 1 presents the result of indexing the article written by professor Jerzy Weres (2010): „*Informatyczny system pozyskiwania danych o geometrii produktów rolniczych na przykładzie ziarniaka kukurydzy*", according to the above, the result is not useful for a person who seeks information in Polish. FAO cooperates also with University of North Carolina, which in its tool Hive Indexer enables selection of thesaurus AGROVOC.

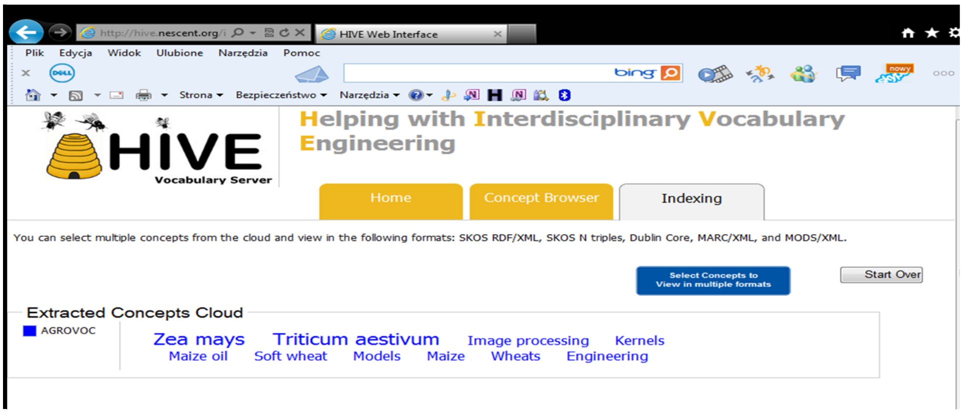


```
- <rdf:RDF xmlns:Tagger="http://agropedialabs.iitk.ac.in/Tagger#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  - <rdf:Description rdf:about="tagger_file12129.pdf">
      <Tagger:agrovoc_tags1>Image processing </Tagger:agrovoc_tags1>
      <Tagger:agrovoc_tags_uri1>http://aims.fao.org/agrovoc-term-info?
        mytermcode=37359</Tagger:agrovoc_tags_uri1>
      <Tagger:agrovoc_tags2> Kernels </Tagger:agrovoc_tags2>
      <Tagger:agrovoc_tags_uri2>http://aims.fao.org/agrovoc-term-info?
        mytermcode=25387</Tagger:agrovoc_tags_uri2>
      <Tagger:agrovoc_tags3> Triticum aestivum </Tagger:agrovoc_tags3>
      <Tagger:agrovoc_tags_uri3>http://aims.fao.org/agrovoc-term-info?
        mytermcode=7951</Tagger:agrovoc_tags_uri3>
      <Tagger:agrovoc_tags4> Engines </Tagger:agrovoc_tags4>
      <Tagger:agrovoc_tags_uri4>http://aims.fao.org/agrovoc-term-info?
        mytermcode=4954</Tagger:agrovoc_tags_uri4>
      <Tagger:agrovoc_tags5> Wheats </Tagger:agrovoc_tags5>
      <Tagger:agrovoc_tags_uri5>http://aims.fao.org/agrovoc-term-info?
        mytermcode=8373</Tagger:agrovoc_tags_uri5>
      <Tagger:agrovoc_tags6> Models </Tagger:agrovoc_tags6>
      <Tagger:agrovoc_tags_uri6>http://aims.fao.org/agrovoc-term-info?
        mytermcode=4881</Tagger:agrovoc_tags_uri6>
      <Tagger:agrovoc_tags7> Wood </Tagger:agrovoc_tags7>
      <Tagger:agrovoc_tags_uri7>http://aims.fao.org/agrovoc-term-info?
        mytermcode=8421</Tagger:agrovoc_tags_uri7>
      <Tagger:agrovoc_tags8> Fruit </Tagger:agrovoc_tags8>
      <Tagger:agrovoc_tags_uri8>http://aims.fao.org/agrovoc-term-info?
        mytermcode=3119</Tagger:agrovoc_tags_uri8>
      <Tagger:agrovoc_tags9> Processing </Tagger:agrovoc_tags9>
      <Tagger:agrovoc_tags_uri9>http://aims.fao.org/agrovoc-term-info?
        mytermcode=6195</Tagger:agrovoc_tags_uri9>
      <Tagger:agrovoc_tags10> Drying </Tagger:agrovoc_tags10>
      <Tagger:agrovoc_tags_uri10>http://aims.fao.org/agrovoc-term-info?
        mytermcode=2402</Tagger:agrovoc_tags_uri10>
    </rdf:Description>
  </rdf:RDF>
```

Figure 1. Example of concepts extracted from article in Polish

This last service is based on KEA tool (*keyphrase extraction algorithm*),which is free of charge and allows indexation of resources towards the thesaurus in SKOS format. The result of indexing is presented on the web site (fig.2). As AgroTagger, Hive-Indexer analyses only concepts from English version of thesaurus AGROVOC, so in practice it concerns the abstract of an article written in English (the results presented in figure 2 concerns the same article written by Jerzy Weres), according to the above, the result is not useful for a person who seeks information in Polish.

Authors, with the use of AgroTagger carried out a row of tests and experiments on various documents in Polish, both of scientific articles and internet publications. A conclusion is explicit, although AGROVOC is a multi-language thesaurus, the indexation process is carried out only in English and in the present form it is is of small usefulness for publication in Polish. The condition for obtaining a reasonable result (but necessary to be translated into Polish) is a good abstract and the set of key words in English.



You can select multiple concepts from the cloud and view in the following formats: SKOS RDF/XML, SKOS N triples, Dublin Core, MARC/XML, and MODS/XML.

Figure 2. Result of indexing article in Polish

A prototype of the indexing system of articles in Polish

During the execution of the project on the knowledge management in the Department of Informatics of the Warsaw University of Life Sciences (N N310 038538 „Narzędzia zarządzania wiedzą w produkcji roślinnej” [Tools of knowledge management in plant production]) a need to index resources has occurred. Since, the use of AgroTagger has not given relevant results, it was decided to create the own system. The main requirement was formulated: indexing papers published in Polish and ultimately profiling indexing to the subject related to agriculture based on AGROVOC thesaurus. Investigating semantic relations between publications was an additional requirement. In order to build a base of varieties of words, free Dictionary of Polish Language with open licenses was used, which was important for authors. This dictionary is being extended all the time. Hobbyists carry for development and editing of the dictionary similarly to editing, which take place in case of Wikipedia, but the content is at a relevant high level. Presently it consists of 17,000 definitions and the whole content may be loaded in the text form. A prototype of the indexing system, was created in the architecture client-server, in the present version of the prototype, PHP language and the base my SQL were applied. Data base of Polish words was designed, which includes inflection forms, data from the dictionary were introduced to this base.

In order to store words and their varieties in the local data base a table was created, the records of which include 5 fields: Identyfikator [Id] – is a main key of the table, Forma Słownikowa [Dictionary Form] – is a basic form of a word, Forma Odmienna [Inflected Form] – is one of forms of a word, Część Mowy [Part of Speech] – includes information on the part of speech, Czy Odmienny [Whether Inflected – includes information whether it is an inflected word. It means that the Dictionary Form occurs in many records, whose number depends on the number of different forms. Table constructed in such a way facilitates finding a word stem, in the column Inflected Form, main search is performed, if a word is in the dictionary then the remaining fields are read out. However, recognition of the part of speech was a serious problem. For this purpose, number of varieties was analysed, which is presented in the diagram (fig. 3). The presented diagram helps automatically to recognize the parts of speech for words placed in the data base. The firs maximum (only one form) includes non-inflected parts of speech, the second maximum (x=6) includes adjectives, the third maximum (x=11) represents nouns, the remaining maxima (x = 23, 38, 57, 77) represent verbs. Unfortunately, these are ambiguous, number of forms of inflection is not sufficient to recognize the part of speech, some adjectives and nouns have the same number of inflections. In order to differentiate adjectives from nouns, the ending of the basic form should be analysed. A similar situation takes place for verbs, but here in practice number of forms of inflections suffices. The following rules help in differentiation of parts of speech. If the basic form ends in "i" or "y" and at the same time number of varieties is bigger than 3 then we deal with an adjective. If the basic form ends in "c" or "ć" and at the same time number of varieties is bigger than 11 then we deal with a verb. If the number of varieties is bigger than 4 and at the same time, the end of the word is different than "i", "y", "c" or "ć" then we deal with a noun.

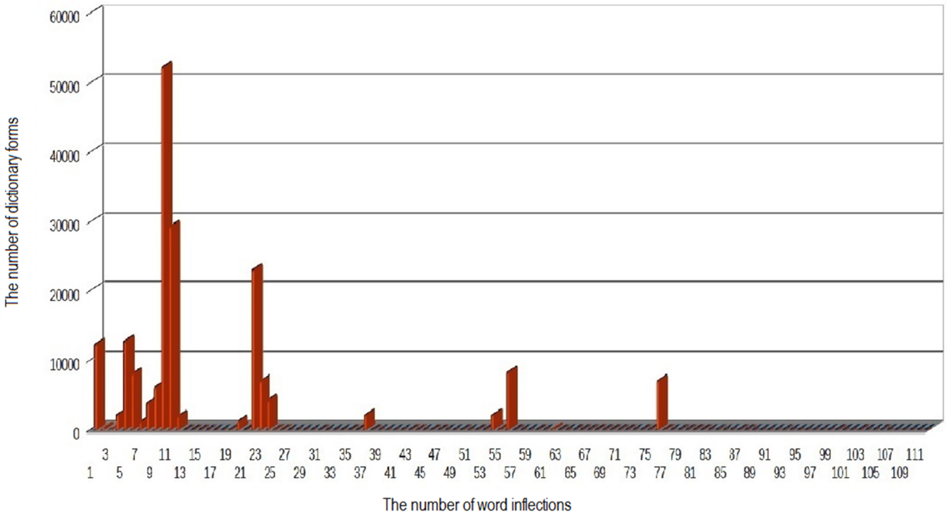


Figure 3. Relationship between dictionary form and the number of inflected forms

During indexing of publications on agriculture, new words in the basic form or inflected appeared. In the specialist language, words, which are not present in the dictionary of Polish language, which we use, occur. Many of them origins in Latin e.g. names of viruses or bacteria (many of them is a combination of two or more words). Principles of inflection of words of foreign origin are generally different than Polish words. Due to complicated inflection of Polish, similarity of new words with those in the dictionary is checked. For this purpose, the method based on Hamming distance is used. When a new word occurs the next word is checked, if it is this next word is not present a new definition is created in the dictionary and then we check in the auxiliary dictionary whether the words which compose the new term have already been recorded. If the word is not recorded, it should be checked whether a stem is already present, if yes, then we add a word as a form of inflection, if not, we add a word to the base as a noun. Indexing in our system consists in determination of the number of times the word appears in the text. In majority of cases, one word corresponds to one concept in the text.

Tests, which have been carried so far, proved that it is possible to divide indexed definitions into two groups: the first group consist in words, which relatively often occur in the text from 5% to approx. 13%, the second group are words which relatively rarely appear in the text. On the investigated group of articles from agrosukces.pl portal, it was determined that papers on the similar subject have the same words which appear in the text the most frequently. The relation between the most popular noun, verb and the second popular noun, which can be used to build a word e.g.: "cow" "have" bacteria" was reported. The gathered data allowed formation of a simple semantic network based on relations between the most frequently occurring words and constittue a basis for additional research trend presented in the paper by Wrzeciono and Karwowski (2013).

Conclusion

The indexing system has accepted so far publications in the text form obtained by copying as a text from web sites or from available scientific articles in doc, pdf, etc. format. The first tests of a prototype allow conclusion that results are promising; in the Polish texts isolation of basic terms is possible. Since, research concern subjects related to agriculture and life sciences, cooperation with thesaurus AGROVOC is a next planned step. The next step should be preparation of interfaces for reading different formats of publications. Ultimately it is necessary to prepare a body stub of texts designated for systematic testing, which would enable further improvement of the system.

References

AgroTagger. Pozyskano z: http://aims.fao.org/agrotagger .

Branny, E. (2005). *Text Summarizing in Polish*. Praca magisterska Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki. AGH Kraków (niepublikowana).

Dolamic, L.; Savoy, J. (2008). Stemming Approaches for East European Languages. *Advances in Multilingual and Multimodal Information Retrieval*, Springer LNCS vol. 5152, 37-44.

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

Karwowski, W. (2010). Ontologies and Agricultural Information Management Standards. *Information systems in managment VI*, ed. P. Jałowiecki & A. Orłowski. WULS Press, Warszawa, 49-56.

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

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

Manning, C.D.; Raghavan, P.; Schuetze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. ISBN: 0521865719.

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

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

Słownik Języka Polskiego. Pozyskano z: http://www.sjp.pl.

Tezaurus AGROVOC. Pozyskano z: http://aims.fao.org/standards/agrovoc/about/.

Weres, J. (2010). Informatyczny system pozyskiwania danych o geometrii produktów rolniczych na przykładzie ziarniaka kukurydzy. *Inżynieria Rolnicza, 7*, 229-236.

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.

AUTOMATYCZNE INDEKSOWANIE ZASOBÓW INFORMACYJNYCH W JĘZYKU POLSKIM DOTYCZĄCYCH ROLNICTWA

Streszczenie. Współcześnie działalność badawcza i produkcyjna wymaga wyszukiwania i gromadzenia różnorodnych informacji, dotyczy to także zagadnień z dziedziny rolnictwa. Obecnie większość zasobów dostępna jest w formie cyfrowej. FAO w ramach portalu Agricultural Information Management Standards prezentuje AgroTagger narzędzie do indeksowania dokumentów z dziedziny rolnictwa, które przeznaczone jest dla języka angielskiego. Ekstrakcja wiedzy jest utrudniona w językach takich jak język polski, posiadających bardzo rozbudowaną fleksję. W języku polskim odmienia się rzeczowniki, czasowniki, przymiotniki oraz zaimki osobowe. Właściwa indeksacja wymaga wstępnej redukcji form fleksyjnych, wobec czego wykorzystano słownik odmian języka polskiego i opracowano program redukujący. Ponadto opracowano i zaimplementowano algorytmy wyznaczania wag odpowiadających ważności terminów uwzględniające częstość występowania terminów i ich pozycję w dokumencie.

Słowa kluczowe: indeksowanie, integrowanie źródeł informacji, sieć semantyczna, zarządzanie wiedzą