Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 437–442
ISBN 978-83-60810-14-9
ISSN 1896-7094

Intermediate information layer. The use of the SKOS ontology to create information about e-resources provided by the public administration

Wojciech Górka, MSc
Research and Development Centre for
Electrical Engineering and Automation in Mining EMAG
ul. Leopolda 31, 40-189 Katowice, Poland
Email: wgorka@emag.pl

Michał Socha, MSc
Research and Development Centre for
Electrical Engineering and Automation in Mining EMAG
ul. Leopolda 31, 40-189 Katowice, Poland
Email: msocha@emag.pl

Adam Piasecki, MSc
Research and Development Centre for
Electrical Engineering and Automation in Mining EMAG
ul. Leopolda 31, 40-189 Katowice, Poland
Email: apiasecki@emag.pl

Jakub Gańko, MA
Institute of Innovations and Information Society
ul. Al. Jerozolimskie 123 A, 02-017 Warszawa, Poland
Email: j.ganko@insi.pl

***Abstract*—Currently, the issue of information search is based on processing a large number of documents, indexing their contents, and then evaluating the level of their adaptation to the question asked by the user. The development of the web allows to offer certain on-line services which make it possible to shop, book tickets or deal with public-administration issues. The objective of the WKUP system (Virtual Consultant of Public Services) is to assist the user in the process of searching and selecting services. The system gives a possibility of natural language communication in the first stage of interaction. This functionality has been achieved by means of the SKOS ontology. The article presents a general outline of the WKUP system architecture and the functioning of the search engine which interprets the user's natural-language questions semantically. The article describes the use of the SKOS ontology in the applied answers searching algorithm.**

I. Introduction

The objective of the WKUP project is to develop a personalized information system which will carry out public administration services with the use of the Virtual Consultant of Public Services (WKUP) based on semantic techniques.

Administration services which can be offered via the Internet are introduced into the public administration step by step, making the citizens' lives easier. These services usually reflect certain procedures and regulations which govern the work of public institutions. On the other hand, the citizen usually uses these services with respect to a larger scale issue he/she wants to settle. In this situation it is necessary to develop a tool which would enable to identify the user's need—the life case, and then guide the user through invoked web services provided by the public administration so that the situation could be dealt with in a complex manner.

The role of WKUP will be to identify the user's issue—life case, give him/her necessary information about that issue, find a relevant public service (or several services), and then to guide the user through the process of complementing the information indispensable to execute the service.

Searching out an adequate process will be done through a preliminary analysis of the user's question and then specifying the issue during the dialogue with the user carried out (to as much extend as possible) in a natural language. At a certain stage of the dialogue, the interaction can be based on selecting the options proposed to the user by the system. During the dialogue the system will collect information about the user (the user's profile will be one of the information sources when user will use system again) and the data necessary to execute the selected process. The user's profile is to facilitate the solution of his/her successive life cases. Both the dialogue and the process of complementing information and the user's profile will be carried out with the use of semantic techniques. The selected service, along with the complemented parameters (those that can be complemented at a preliminary stage of service execution and on the basis of legal and administration terms of the service) will be invoked in a government electronic system.

The architecture will consist of four functional layers (Fig. 1). Two layers, i.e.: the natural language processing layer and knowledge layer, will refer to the WKUP user interface functionality, while the processes layer and web services layer are related to the functional range of the Semantic Broker.

The natural language processing layer will consist of two modules—chatterbot and semantic search engine based on the SKOS ontology. The role of the chatterbot will be to make conversation with the user about casual issues (weather etc.) The role of the semantic search engine will be to find out what the users' need is.

The following chapters describe a part of the system related to the semantic search engine.

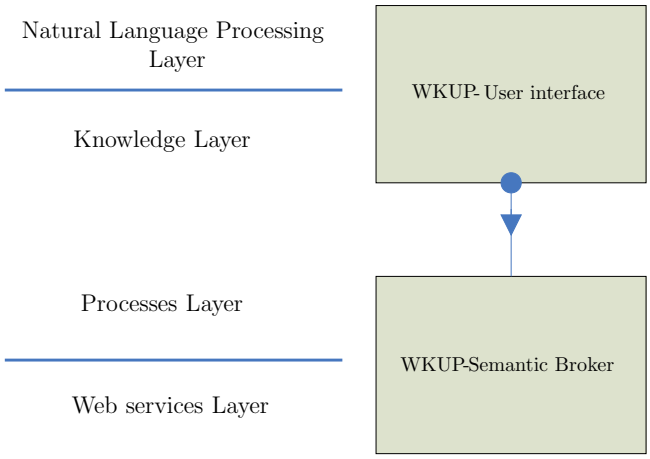

Fig. 1. WKUP architecture

II. General outline of search issues

There are many solutions in the realm of information search which allow to index the information contents and search for documents based on the contents. The full-text search solutions are mostly based on statistics and there have been many algorithms developed in order to standardize the search results [1]. A relatively new solution are algorithms which allow to cluster the search results [2]. Clusterization introduces the documents selection with respect to areas of interest (a sort of categorization) based on words used in a given text. A category is, to certain extent, a representation of the document contents determined on the basis of the statistics of words used in the document. Examples of such solutions are Vivismo [3] and Carrot2 [4].

One of the full-text search products is the Lucene search software [5]. The software enables to create a properly compressed index and to efficiently search for documents (even concrete places in documents) which are the answer to the question asked by the user. Additonally, Lucene makes it possible to create adapters which allow to browse different types of documents (Microsoft Office documents, XML documents, PDF documents, etc.)

The accuracy of the information search is achieved due to the use of semantic webs solutions [6]. Semantic webs allow to describe information in a formal way and to introduce interdependencies between particular pieces of information. This way the information search is broader. The use of semantic webs will allow the search tools developers to design new–quality products. The search tools, equipped with the knowledge about the concepts hierarchy and their interdependencies, will make an impression of intelligent software. Such knowledge allows to search not only for the key words given by the user but also for the related concepts, and shows how this relation is made. The example are synonims of terms given by the user, semantic relations whole-part or other relations between words.

Irrespective of the development of information technologies there are works caried out in the realm of text corpuses[1], which enable to determine, among others, dependencies between words and the frequency of their occurrence in texts [7]. Such works allow to create word nets (WordNet [8]). The works on the word net for the English language have been carried out since 1985. The works on other European languages (Czech, Danish, German, Spanish, Italian, French, Estonian) were carried out between 1996-1999 within the EuroWordNet project [9]. In Poland the works have been conducted within the plWordNet project [10]. Constructing a word net is done automatically, to a certain extent, thanks to the use of the Polish text corpus. The data from word nets, actually—relations between words, can be used to associate the words which appear in the indexed texts. This way it is possible for the user to find documents on the basis of the question in which the key words included in the document have not been used directly. Thus this solution is similar to proposals derived from the semantic webs concept.

In the realm of information search it is possible to determine the qualities of systems whose objective is to answer the questions. An example is the AnswerBus system [11] based on the knowledge indexed by Internet search tools. The search results are interpreted in an adequate way so that the information looked for by the user could be extracted from the document found by the search tool.

Another interesting solution is the PowerSet search tool [12]. The objective of the tool is to answer the user's questions on the basis of resources in the Wikipedia service. The tool operates on the basis of structures which enable to determine the question context, to select answers into certain thematic categories, and to find related concepts.

III. Motivation of the execution

Full-text search is based on the statistical analysis of words included in the processed documents. The works "An Introduction to Information Retrieval" [2] and "Term weighting approaches in automatic text retrieval" [1] present different issues related to full-text search tools operations (indexing, compressing the index, analysis of the user's question and the answer to this question). It is worth mentioning that this approach to searching is based on statistical methods and requires plenty of data in order to achieve accurate and appropriate results. A large number of words in a document, as well as a large number of documents, allow to better select the words which are characteristic of the given document—key words. The solutions based on full-text search tools achieve better results in the case of a large number of long-text documents.

[1] A large and structured set of texts (now usually electronically stored and processed). They are used to do statistical analysis, checking occurrences or validating linguistic rules on a specific universe. They are the main knowledge base in corpus linguistics. The analysis and processing of various types of corpuses are also the subject of much work in computational linguistics, speech recognition and machine translation. (*source: Wikipedia*)

Within the WKUP project a different solution was applied.

As it was mentioned before, the objective of the information system which is currently being developed is, among others, to give the user advice on the life case described by the user. Thus the user will ask a question and the system will propose one or a few possible pieces of advice (previously defined in the system). The advice will have a form of short descriptions explaining the operations the user will have to do in order to solve his/her issue (the descriptions have to be readably short and understandable to the user). For example, if the user asks for help because he/she has a stomachache, the advice given should be a message advising the user to see a GP and proposing an appointment via the WKUP system.

The solution can be briefly described as a set of a large number of answers to potential questions of the users - similarly to FAQ lists (Frequently Asked Questions). A potentially large number of the pieces of advice is the encouragement to develop a search tool which proposes a piece of advice (answer) best related to the real-life situation (question).

A small number of words in the text (short contents of the advice) has a negative impact on the efficiency of the document indexing process. The algorithm calculating potential key words for a given document may take into account wrong words due to limited size of the text.

It is also possible to assume that the potential questions range is, to a certain extent, determined. For example, if the system provides information from the field of medicine, the questions asked to the system will be related to illnesses, symptoms or advice connected with the organization of the national health system.

Additionally, it is important to notice that the users who ask questions will not necessarily use the words and terms included in the advice. Associations between the terms used in the question and in the advice may be even more distant than previously described terms associations in semantic webs. The application of semantic techniques will allow the user to use casual words to form the questions which are asked to the domain system comprising specialized vocabulary. The semantic technique allows an average skilled user to make use of the domain system.

Two presented reasons: small size of questions and answers, as well as different ranges of vocabulary used by the user and the information system, are the basis for the solution which allows to interpret the users' questions and to control the results displayed by the search tool.

IV. Ontologies and SKOS

In order to present the solution we will focus on issues related to technologies applied in the solution development.

Ontology is a branch of philosophy which tries to describe the structure of reality. As understood by philosophy, ontology allows to explain relations between entities, qualities of these entities, etc., so that the reality could be described. In order to "understand" a section of reality, a computer needs the data that describe this reality, i.e. ontologies. The ontologies (as understood by the information technology) and their application are within the interests of the World Wide Web Consortium (W3C). In 1997 a standard was proposed, and as early as in 1999 W3C published the Resource Description Framework (RDF) standard [13]. The standard was complemented in 2004 with the RDF Schema (RDF-S) specification [14].

RDF allows to record triples of concepts. Each triple is a subject-predicate-object expression. Such a way of concepts recording forms a network of definitions (each object can be a subject in a different triple). Fig. 2 features a sample ontology diagram on which the concepts (circles and squares) are depicted along with their relations (arrows with names).

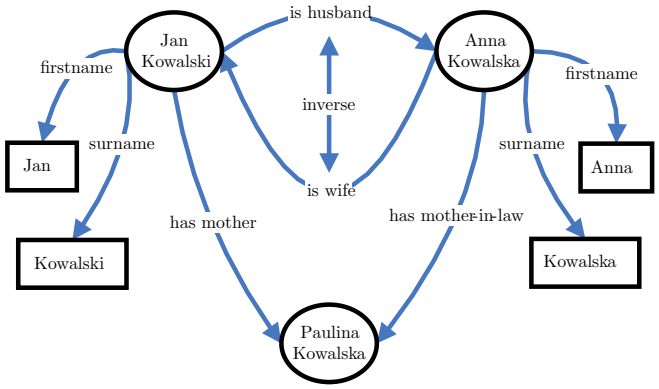

Fig. 2. Sample ontology

RDF-S introduced the possibility to build meta-concepts: classes, sub-classes, features. It also launches a non-standard way of defining the name of the notion (*label*) and its description (*comment*).

The next stage to extend the semantic web standards was to increase the expressiveness of languages intended for ontology recording. W3C published the OWL (Web Ontology Language) standard [15]. The language allows, among others, to express the number of concept sets, to show how one concept belongs to or differs from the other, to identify necessary and sufficient conditions for a given concept. Greater expressiveness of the language allows to verify concepts added to the ontology and to search out certain facts and features indirectly. Additionally, OWL makes it possible to integrate two ontologies by means of associating their identical concepts.

Therefore, ontology description standards allow to describe concepts and the network of links between concepts.

The SKOS specification (Simple Knowledge Organization System) [16], developed and extended under the auspices of W3C, defines an ontology which allows to express the basic structure and contents of concept diagrams, including thesauruses, thematic lists, heading lists, taxonomies, terminologies, glossaries, and other kinds of controlled dictionaries. The specification is divided into three parts: SKOS-Core [17] [18], SKOS-Mapping [19] and SKOS-Extensions [20].

SKOS-Core defines basic concepts and relations which enable to develop concepts and relations between them. SKOS-Mapping introduces relations which allow to describe

similaries between concepts created in different ontologies. SKOS-Extensions introduces extentions of the intensity of hierarchical relations from SKOS-Core.

The SKOS ontology assumes that concepts are described by elements linked by means of the *subClassOf* relation with the *Concept* element.

Each concept can be labelled. The SKOS ontology extends the labels that can be used:

- *prefLabel* (chief label of a given concept)
- *altLabel* (auxiliary label, alternative for a given concept)
- *hiddenLabel* (hidden label, e.g. for casual words or other words treated as "hidden" due to other reasons).

The concepts can be linked into hierarchies by means of *broader* and *narrower* relations. The SKOS-Extensions specification introduces extra semantics of hierarchy relations, among others by the following relations:

- *broaderInstantive*/*narrowerInstantive* (express context hierarchies—instances, e.g. Dog and Azorek[2]).
- *relatedPartOf*/*relatedHasPart* (express the whole-part semantics, e.g. Car and Wheel).

The SKOS ontology also provides the class definition which describes a set of concepts—*Collection*. Such a set can help to manage the ontology and facilitate its edition by grouping concepts of similar meanings. Possible ways to use the structures of concepts built on the basis of the SKOS ontology were described in use cases [21]. What is derived from these use cases is, among others, the application of SKOS to the following:

- to order and formalize the concepts used in a given domain, to search—on the basis on the concepts and a part of relations between them—for resources assigned to the concepts,
- to search for information in different languages (thanks to an easy method of translating labels in the ontology with an unchanged relation structure),
- to label press articles, TV programmes, etc. with key words from a thesaurus recorded in accordance with the SKOS ontology.

The above objectives of the SKOS ontology satisfy, to a high degree, the requirements of the search tool in the WKUP system. Therefore a decision was made to apply this ontology. The application was justified by the possibility to provide the tool with a wide and, at the same time, precise "understanding" of concepts. Thanks to semantics it is possible to record the relations between concepts which, in turn, allows to better interpret the questions.

V. Performance method

The use of the SKOS ontology in the WKUP system consists of two stages: edition and production (search tool operations). Fig. 3 presents the way of using the concepts, defined in accordance with the SKOS ontology, with a view to search for certain resourses—data—related to these concepts.

[2]Popular dog name in Poland.

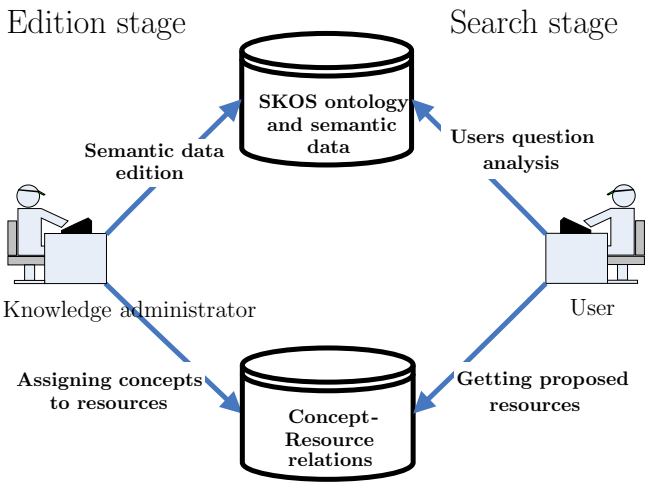

Fig. 3. The use of concepts defined in accordance with the SKOS ontology in the search process

At the edition state (before the system starts) the administrator defines concepts and their mutual relations. Then he/she creates relations of the defined concepts with the data which are to be searched for. The ontologies defined in this manner are used at the search stage (production operations of the system). The user's question is analyzed based on the used concepts. The identified concepts are processed. On the basis of mutual relations between concepts, the best fitting answers of the system are found—the resources the user is looking for.

The analysis algorithm of the user's question was divided into successive stages. The first stage is "cleaning" the user's question from redundant non-alphanumeric signs as well as lemmatization of particular words in the sentence. For the statement prepared in such a way, at the next stage the best-fit concepts are searched for based on their labels (relations *prefLabel*, *altLabel* and *hiddenLabel*). In the case when the found concepts are not related to the resources, the relations *broaderInstantive*, *broader* and *relatedPartOf* are used in order to search the web for the concepts which have certain resources assigned. This allows to find the concepts whose meaning is broader than the meaning of the concepts used in the sentence.

The *related* relation is treated in a special way. Thanks to the *related* relation, several concepts which lead to the same resource make the resource "stronger" by assigning a higher searching priority to it. This way it is possible to model the relations between concepts derived from the knowledge about the specifics of the given domain for which the concepts are modelled. The last stage of the sentence analysis is the use of information about the words location with respect to one another in the user's sentence. The words which are closer to one another and point at the same resource simultaneously raise the priority of the found resource. This results from the prerequisite that, usually, the words which determine the same object are located close to one another in the sentence.

Such analysis allows to present the found resources to the user according to the assigned search ranking.

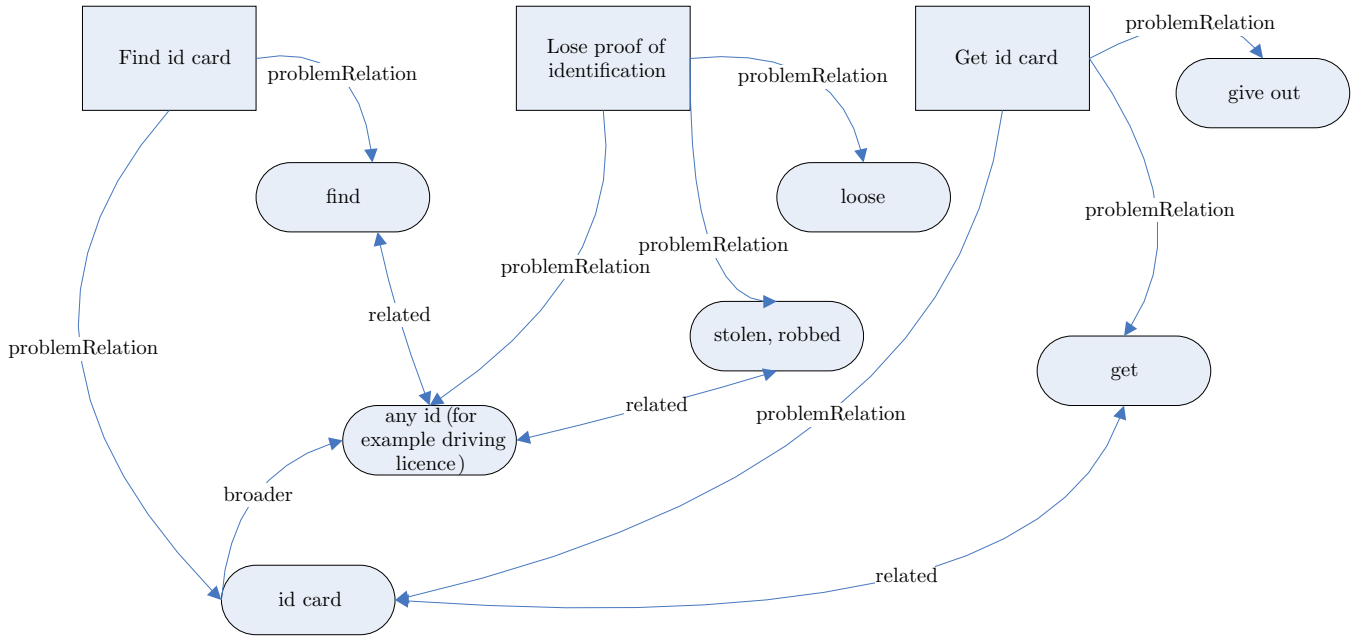

Fig. 4. Sample SKOS structure and its relation to the resources to be searched for

Fig. 4 features a sample SKOS concepts structure and its relation to resources that are to be searched for. Three issues (real life situations) have been defined: finding an ID, losing an ID and getting a new ID. Additionally, the following concepts have been defined: finding, loss, theft, getting and issuing. The *related* relations allow to "strengthen" certain relations other than *broader* and *relatedPartOf*.

Building a net of concepts and assigning resources to the concepts allow to model the system answers to the user's questions. This way the data administrator, who defines the system answer by himself/herself, has a clear picture of the system behaviour with respect to a given class of questions. Such a solution is more deterministic than full-text search tools which operate on the basis of statistical methods only. Additionally, to improve the data administratord's operations in the system, the mechanisms were introduced which function in traditional search tools solutions, but at the edition stage of the ontology. Thus the possibility of automatic collection of concepts from the indexed elements (descriptions of life cases) was applied, and.the process of assigning the concepts to life cases was automatized. In order to perform this task, the algorithm was used to calculate normalized words priorities for documents (*dt* indicator) [2]. The algorithm allows to calculate the adequacy ranking of a given word for the indicated life case. Therefore the work with the tool can start from automatic indexing of life cases and then proceed to successive introduction of revisions by means of successive introduction of relations between concepts, changing labels and their classification (pref, alt, hidden), etc.

VI. Conclusions

The presented solution is a proposal to solve a certain issue related to information search. It seems that the solution can improve the search in resources which are limited in terms of the number of indexed documents, and in the situation in which it is assumed that the users will ask "questions" to the search tool. The solution appears especially adequate in the case of the so called FAQ lists. They define ready answers to certain questions and, more importantly, the questions are usually relatively short. In such cases full-text search tools can have problems to properly index the contents.

The solution is at the prototype stage now and its operations have not been checked in practice yet. On the basis of the conducted tests it seems, however, that the efficiency of the search tool operations depends mainly on a well constructed ontology. Therefore the ontology is the key element which affects the functioning of the system.

Practical results of the search tool operations and the drawn conclusions will be the topic of the next publication.

References

[1] Salton G., Buckley C. , "Term weighting approaches in automatic text retrieval. Information Processing and Management 32", pp. 431–443. Technical Report TR87-881, Department of Computer Science, Cornell University, 1987
[2] Manning C.D., Raghavan P., Schütze H., "An Introduction to Information Retrieval," Cambridge UP, Draft of July 1, 2007
[3] Vivismo, http://vivisimo.com
[4] Carrot2, http://www.carrot2.org
[5] Apache Lucene, http://lucene.apache.org, http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/20040504/
[6] Semantic Web, http://www.w3.org/2001/sw/
[7] Przepiórkowski A., "The Potential of The IPI PAN Corpus", Institute of Computer Science, Polish Academy of Science, Warsaw
[8] WordNet, http://wordnet.princeton.edu

[9] EuroWordNet, http://www.illc.uva.nl/EuroWordNet
[10] Polski WordNet, http://www.plwordnet.pwr.wroc.pl/main
[11] AnswerBus, http://www.answerbus.com/index.shtml
[12] PowerSet, http://www.powerset.com
[13] W3C, Resource Description Framework, http://www.w3.org/RDF
[14] W3C, Resource Description Framework Schema, http://www.w3.org/TR/rdf-schema
[15] W3C, OWL Web Ontology Language, http://www.w3.org/TR/owl-features
[16] SKOS, Simple Knowledge Organisation System, http://www.w3.org/2004/02/skos
[17] SKOS Core Guide, http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102
[18] SKOS Core, http://www.w3.org/2004/02/skos/core.rdf
[19] SKOS Mapping, http://www.w3.org/2004/02/skos/mapping.rdf
[20] SKOS Extensions, http://www.w3.org/2004/02/skos/extensions.rdf
[21] SKOS UseCase, http://www.w3.org/TR/2007/WD-skos-ucr-20070516/