Semantic Web: Asking the Right Questions

Włodzisław Duch
Department of Informatics, Nicolaus Copernicus University, Toruń, Poland.
Google: W. Duch

Julian Szymański
Department of Electronic, Telecommunication and Informatics
Gdańsk University of Technology, Gdańsk, Poland
Email: julian.szymanski@eti.pg.gda.pl

Abstract: Searching for information on the Internet or in free texts requires knowledge of proper keywords. Some information is quite difficult to find and in most cases many irrelevant documents are retrieved. The same problem is evident when large help systems are used. Techniques based on clusterization and visualization of information have been used to alleviate this problem, but in practice such methods are not very useful. A novel approach based on asking minimum number of questions that should make the query unique is presented here. The background knowledge stored in semantic memory has been collected using WordNet and many other linguistic resources. A simple experiment in medical domain demonstrates the potential of this approach.

Keywords: Semantic Web, semantic memory, information retrieval, knowledge acquisition, Wordnet, NLP.

I. Introduction

Internet and large electronic databases completely changed the way information is searched, maintained and acquired. Yet instead of boosting productivity increasingly large proportion of time is spent on searching. With well over 10 billion pages indexed by each of the major search engines the problem is getting worse every year. In addition the overall quality of information in the Internet is degrading, with large number of spam pages containing advertisement and misleading information. The dream of semantic Internet, discussed since the 1998 seminal paper of Tim Berners-Lee "Semantic Web Road map"[1] (see also [1][2]) is still far from its realization. Semantic web promised to add definitions or the meaning to text data by providing metadata. The World Wide Web consortium (W3C) concentrated on technologies, such as the XML (eXtended Markup Language) and the RDF (Resource Description Framework), a markup language for describing information and resources on the web. In principle these technologies allow for semantic description of texts, and the use of markup tags in the search queries. However, adding such tags manually proved to be too tedious and therefore only a few interesting semantic web projects have been created so far. As a result Semantic Web technologies grow much slower than expected. Tools for automatic or semi-automatic tagging of standard HTML documents are still rare, although performing partial tagging, adding for example named entity tags such as company or person's names, is not difficult, but already would be quite useful. The goal of creating semantic Internet, with easily accessible and understandable information, proved to be more difficult to reach then initially thought.

There is a clear need for algorithms that will help to increase relevance and retrieval accuracy of information search without relying on special tagging. There are several ways this problem may be approached, reviewed in the next sections. Most queries are ambiguous, so precisiation of the user queries, that is asking minimum number of questions needed to define the subject, is an interesting possibility that so far has not been explored. In section four our approach to this problem will be presented and in section five an experiment with psychiatric diagnosis is described.

II. Semantic Web and Information Retrieval

To outline the background for our novel approach to the problem of information retrieval first some remarks on the efforts to use semantic web technologies for information management are made, followed by review of other technologies that may be used to improve retrieval of relevant information.

RDF metadata model is relatively simple, based on triples encoding subject-predicate-object. These triples express relationships between subjects and objects, for example <article, author, person>. The information is written in a formal way that computers may understand, and special care should be taken to remove ambiguity, as words such as "article" have many meanings (this word has at least 5). This ambiguity is quite difficult to handle, and in fact is an indication of the general problem in natural language processing (NLP), the lack of precise meanings of words. The context provided by RDF is simply not sufficient for unique interpretation. A number of specialized RDF schemas have been defined with the help of the RDF Schema vocabulary description language (RDFS), or richer ontology languages such as the DARPA's Agent Markup Language (DAML) together with the Ontology Inference Layer (OIL), defining logical system to make automatic inferences and to assist with integration

[1] http://www.w3.org/DesignIssues/Semantic.html

of different and incompatible ontologies. More recently Web Ontology Language (OWL) specification for authoring ontologies has been endorsed by the W3C organization as the family of knowledge representation languages based on description logics. The RDF query language (called SPARQL), that has been standardized only in January 2008, should contribute to further development of semantic web. Really Simple Syndication (RSS) language has gained a lot of popularity, allowing to syndicate the site content.

These and many other technologies have been used to define RDF schemas that may be searched by software agents in real-time to disambiguate RDF data. Schemas should provide semantics needed to extract the meaning from the RDF triples, thus forming a critical layer of the Semantic Network. The SchemaWeb[2] is a directory of RDF schemas that have been manually created to describe specific ontologies. For example, the beer ontology models types of beer and brewers/brands, wine ontology does the same for wine, ontologies exist for some types of music as well as several branches of science. The Agricultural Metadata Element Set defines agricultural information standards and ontologies that may be used for effective data exchange, interoperability and resource discovery, the Semantic Web for Earth and Environmental Terminology (SWEET) project provides a common semantic framework for various Earth science initiatives, and the Common European Research Information Format (CERIF) is an effort of European Commission to represent metadata about current scientific research.

One area where semantic web technologies have made a big impact is in the geographical information systems (GIS), involving Global Positioning Systems (GPS) in cars and portable devices, Google maps, Wikitravel and other systems that benefit from precise geospatial information about objects stored in their catalogues. The Geonames Ontology is a large-scale project with over 6 million toponyms (names of topographical objects) that have the RDF web references and a unique URL. The Geonames website[3] allows to search such objects like rivers, cities, landmarks, temples or hotels, showing them on the map and automatically linking them to sources of additional information. Relations between toponyms are described by other services. Another quickly growing area is in the social networks that already have tens of millions of users. The Friend of a Friend (FOAF) project[4] is creating a Web of machine-readable pages describing people, their interest, recommendations, schools they have attended and organizations they belong to, creating links between them. These links may automatically be extended, helping to discover new interesting connections. The popularity of blogging has also created a demand for automatic integration of knowledge sources, in particular web syndication and the Traceback[5] techniques.

A few successful semantic web technologies have also been deployed in science, facilitating new discoveries. Internet has opened the door to vast amount of information contained in research papers and specialized databases. The problems in life sciences, molecular medicine, pharmacology or neuroscience are so complex that without tools to associate and integrate data from distributed knowledge sources progress will be very limited. The current mean time from discovery of a gene mutation to introduction of effective drugs is 17 years. Science Commons[6] is an MIT initiative aimed at accelerating scientific discoveries through a new web-enabled collaborative infrastructure that should help to reuse results of research, providing access to experimental datasets for further analysis, enabling fast access to research materials and integration of data from heterogeneous distributed information sources.

As a proof of the concept Neurocommons[7], an open source knowledge management platform for neuroscience research is being created. A lot of knowledge is contained in the open biomedical abstracts and articles published in open access journals. Natural Language Processing, and especially text mining techniques, have been used to organize and structure this knowledge. Software for data analysis, with particular focus on neurodegenerative diseases, will also be provided by Science Commons. Leading pharmaceutical companies, including Pfizer, Biogen and Novarities, are involved in this project, helping to extract, from unstructured text across different subdomains, implicit semantics, common terms and relations between them. An open knowledgebase of RDF annotations of biomedical abstracts and major neuroscience databases is going to be converted into an annotation graph. It may be accessed using the SPARQL query language. For example, searching for biological processes that may lead to "neural dendrite degeneration" gene ontologies are consulted and concepts such as "dendrite morphogenesis", development, regeneration and regulation are retrieved, genes that are involved in these processes identified and links to relevant research papers provided. Although these tools are still not easy to use they save a lot of time for literature search, tracing the links between numerous sources. The size of such undertakings is enormous, with existing bases reaching hundreds of millions of RDF triples and still covering only a small fraction of biological knowledge.

International Neuroinformatics Coordinating Facility (INCF)[8] (currently based in Sweden) is concerned with development of infrastructure for neuroscience research, but many of its goals depend very much on the semantic web techniques, integration and management of text data and other symbolic information. Several projects in neuroinformatics are aimed at understanding of specific brain subsystems, for example the Visome project that is mostly focused on color vision [5], or the cerebellum project, trying to inte-

[2] http://www.schemaweb.info
[3] http://www.geonames.org
[4] http://www.foaf-project.org
[5] http://en.wikipedia.org/wiki/TrackBack
[6] http://sciencecommons.org/projects/data
[7] http://neurocommons.org
[8] http://incf.org

grate experimental databases, software for data analysis, models of neural processes and relevant literature.

Full understanding of animal (including human) behavior in all its complexity requires description of the environmental impact on organism leading to the regulation of gene expressions in the inherited genome. The BeeSpace[9] project tries to reach the understanding of social behavior of honey bees on the whole genome scale. Each honey bee lives in a complex environment, in a society that divides labor depending on the age and genetic disposition of its members. Data from microarray gene expression experiments for hundreds of bees that play specific societal roles is collected in an interactive system, and linked to literature on insect behavior. The BeeSpace Concept Navigator software uses statistical text mining analysis facilitating development and testing of hypothesis about functional relationships between genes and behavior, navigating users through diverse databases and literature sources.

Common themes in the semantic web approach include content-based, rather they keyword-based information retrieval, annotation of unstructured information using metadata, treating the whole web as a large database that should be readable by intelligent software agents, providing various services that require now a lot of human effort.

Semantic web technologies have already found a number of interesting applications, but are still rather difficult to use and rely on a lot of manual work to create ontologies.

III. Statistical NLP

Not all technologies depend on RDF and semantic net ideas. While projects described above are likely to succeed in their specialized, narrow domains, it will be hard to scale them up to arbitrary domains and use them in search engines for information retrieval. Common sense ontologies are still missing and tools for (semi)automatic annotation of unstructured texts have yet to be developed. Large collections of static web links to interesting sites are created by social networks, for example Del.icio.us has over 3 million users and 100 million bookmarked URLs. Some web links may be replaced by proper keywords that help to trace web pages even when the URLs are changed (Google and other search engines allow for embedding queries in html). However, many web pages are difficult to find through the search engines, either because they either use common words, or untypical words that users will not think of making queries. A simple method for creating links that are easily traced if the URL is changed is to use unique words in the meta-tags, and rely on the search engine instead of giving direct link. For example, adding a new unique keyword combining parts of the first and the last name (wloduch) will match only a single document (or a group of your documents) after indexing by the search engine, so the process may be made quite transparent to the user.

Semantic search may also be based on statistical approach to NLP that uses large text corpora to learn the meaning from the context. The Interspace project[10] [6], one of the DARPA's Information Management Program projects, has been one of the largest efforts to develop a prototype environment for semantic indexing of information based on statistical clustering of concepts and categories. The goal was to enable scalable information retrieval with semantic interoperability across many subject domains [7]. The project used Concept Space thesauri, develop for digital libraries, based on a hybrid symbolic/numeric representation of concepts. This allows for visual representation of relationships between concepts and leads the user to new keywords presenting similar concepts that may be used to enhance the search. Tests were conducted with document collections that include engineering and medical literature. In particular the MedSpace project [8] has been used as a large-scale testbed for clinical medicine. Semantic indexing of over 9 million MedLine abstracts extracted over 270 million noun phrases, of which 45 million were unique. Perhaps because of the size of these projects they ended up as prototypes and are not used in practice.

Clustering techniques have been popular in information retrieval in the recent past. The ideas is to arrange the search results into groups that show some similarity, instead of using only relevance to the search query. This approach may help to focus on desired subgroups of results that may otherwise be missed as less relevant. This is especially helpful during initial exploration of the topic, when ambiguous keywords are used instead of more specific ones. Some of the early search engines using this technique include WiseNut (now removed), Teoma (now replaced by Ask.com) and Vivisimo (now a part of Clusty.com). For example, searching for "Virus" the Ask.com will propose the following categories: Virus Disease; Computer Virus; Human Virus; Virus Information; Type Viruses; Virus Info; Free Virus Scan; Virus Encyclopedia; West Nile Virus; Hoax Viruses. Not all these categories are at the same level of generality, as the search engine is not using hierarchical ontology but rather statistical correlations. It also proposes to expand the search by switching to related keywords: Bacteria; Free Antivirus; Fungi; Virology; HIV; AIDS; Chicken Pox; Bacteriophage; Influenza; Flu. Searching for "Word" will show such clusters as "Microsoft Word", and "Word Games". Others search engines that do clustering include iBoogie, Killerinfo, Carrot2 and Kartoo. Clustering texts without knowledge is rather difficult, therefore these search engines do not work too well, creating many clusters for the same categories, for example Hoax viruses and Free Virus Scan are Computer Viruses, or showing similar documents in different clusters. Adding some ontology-based knowledge would certainly improve results of clusterization-based search.

Clusterization may also be used for visualization of texts, showing maps with peaks corresponding to similar docu-

[9] http://beespace.cs.uiuc.edu

[10] http://www.canis.uiuc.edu/interspace/

ments that may be labeled by the relative frequency of keywords. Several visualization technologies have been invented in the Pacific Northwest Laboratory (PNL)[11], but the only such technology that has been used in larger scale experiments seems to be WebSOM [9]. The method is based on Kohonen's Self-Organizing Map clustering and visualization algorithm, and is used in the SOMLib Digital Library Project[12] to provide automatically spatial organization of documents by their content. However, it remains to be seen how successful this attempt will be, as a number of web services gave up this idea after a short period of experimentation (including Cartia and a few astronomical journals that have displayed catalogues and papers using WebSOM).

Computation with perceptions and with information described in natural language has been called by Lotfi Zadeh "a new frontier in computation". His approach [10], based on fuzzy concepts for handling uncertainty, is aimed at the precisiation of meaning by restricting natural language to a subset that allows to formulate precise definitions of new concepts. Although these ideas look attractive there are no examples of useful applications for concept disambiguation or information retrieval.

The Semantic Web ideas have been invented by computer scientists and experts in logic, stressing the need to create machine-readable information, ignoring cognitive aspects in favor of purely formal methods. The key issue is disambiguation of information and finding associations that are interesting for the user. Neurocognitive inspirations may help to solve this problem in a natural way. So far only human brains are capable of understanding complex information structures contained in texts. Although we have a rough understanding of this process [11][12] there are no good approximations that do something similar in software systems. Knowledge representation schemes used in artificial intelligence are quite crude reflection of the spreading activation processes in biological neural networks, encoding salient features of concepts. In cognitive psychology different types of memories are distinguished. Although the RDF triples capture some relations between concepts and features each concept is in fact involved in many relations of different types. Initial activation of neural subnetwork coding phonological representation of the concept is spread to associated concepts, depending on the current context. Ontologies do not capture the dynamics of this process. Ultimately human level language and information management competence may require algorithmic implementation of similar mechanisms. Semantic network models [13]-[15] encoding relations between objects and their features reflect to some degree biological processes, but large-scale networks containing knowledge describing common concepts needed for general information retrieval applications do not exist. It is our goal to enhance Wordnet [16] and other lexical sources with vector representations [17]-[20] that approximate initial stages of the spreading activation processes, allowing for correct associations and semantic disambiguation of concepts. Basic meanings of concepts describing real objects are stored in semantic memory [11], while current context is stored in the episodic memory [12]. The most popular approach to the semantic memory (SM) models is based on the Collins and Quillian hierarchical model [14], Collins and Loftus model of spreading activation [15], and the feature comparison model [12]. Such knowledge is sufficient not only for disambiguation of concepts in the current context but also to formulate questions in word games [18].

The progress in information retrieval due to the semantic web and statistical NLP methods has been rather slow, showing the difficulty of dealing with the natural language in unrestricted domains. In the next section a novel approach based on precisiation of queries is presented. The main idea behind it is to suggest additional keywords by asking the user a minimum number of questions that should be sufficient to precisely identify the subject of the query.

IV. Retrieving Information by Asking Questions

Asking questions requires background knowledge. SM may be used in natural language dialogue systems, word games, formulation of questions, precisiation of queries for semantic search, and many other applications [17]. Biological processes behind it involve spreading activation in neural networks, and are thus difficult to approximate due to their dynamical nature. Each concept is a short-lasting activation of a particular configuration, with phonological as well as extended components (depending on the type of concept they may involve activation of visual, auditory or motor cortex [20]), and the main channels of spreading activation are between the concept and features that are applicable for concept description. In computer algorithms this may be replaced by typed relations (predicates) between concepts (objects) and keywords (features), with weight values expressing the strength of such relations. This leads to a slight generalization of the RDF triples [4], with added weight values. Such structures are called bellow wCRK (weighed Concept - Relation Type - Keyword), and are used to form components that describe concepts. An example of wCRK is shown in Fig 1.

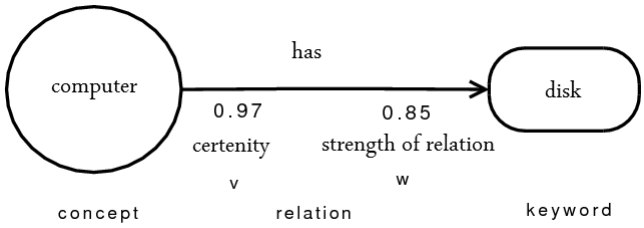

Fig 1. Simple sentence „computer has a disk" coded using wCRK knowledge representation.

[11] http://infoviz.pnl.gov

[12] http://www.ifs.tuwien.ac.at/~andi/somlib/

The triplet: concept – relation – keyword, forms a "word layer" of the system. The v weight allows for coding the uncertainty of knowledge, the w value codes the strength and direction of the relation, indicating how typical the feature is for a given concept. The wCRK shown in the Fig 1 says that with high confidence (v = 0.95) computer has a disk (most do, w = 0.85).

Automatic creation of concept descriptions using RDF representation is the main problem of semantic web, and the same is true in case of semantic memories. Our SM system has been based on the WordNet [16] and several other lexical resources [17], implemented using Microsoft .NET technology, with the Microsoft IIS as an application server. It facilitates system access from heterogenic environments, WWW and multimedia interfaces, web services, enabling cooperative semantic memory usage. A set of libraries encapsulating communication with database and realizing numerical calculations on the semantic space forms the business intelligence layer, called semAPI (Fig. 2). Such approach facilitates scalability, clear functional borders and makes complex data operations easy, providing uniform database access for different applications.

Presence of numerous many-to-many relations poses high demands on system processing. A relational database offers standard and widely accepted mechanisms to deal with this issue. Microsoft SQLServer was used here as a container for conceptual information storage. It serves as a repository for holding wCRK structures, with weighted relations that allow for modification and learning of new knowledge. The architecture of the system is depicted in Fig 2.

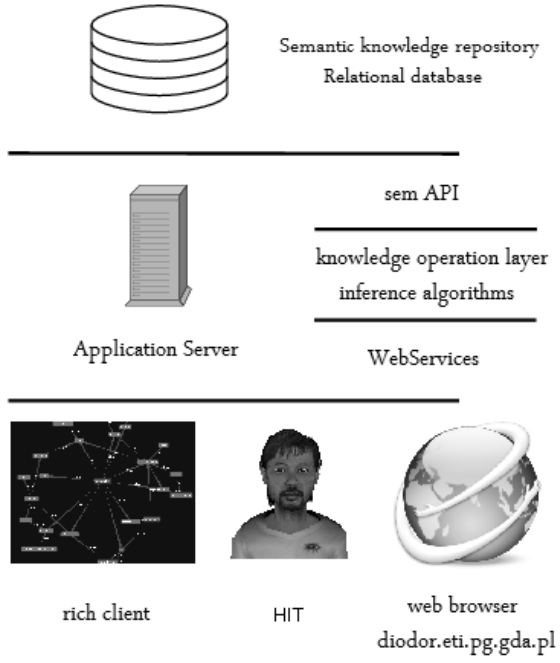

Fig 2. Architecture of the semantic memory system

To avoid direct operations on the database, requiring tedious programming, an intermediate translation layer has been implemented, making this process flexible and programmer-friendly. The role of "business intelligence layer" (semAPI in Fig. 2) is to map database tuples on the application objects, making the interaction with database transparent for programmers. It also ensures realization of atomic functions for building and operating large logical structures, such as classes for numerical calculations performed on semantic space, providing fast algorithm for turning ontology-oriented semantic representations into their numerical representations (see below).

The semantic memory system is accessible through several interfaces that serve as gateways for users for retrieval or entering of new data.

a. Semantic space visualization

Data stored in semantic memory in the wCRK form are hard to analyze directly. To make this work easier a TouchGraph [21] component has been used as a graphical tool for visualization of concepts and their relations. A java applet working with semantic memory web services creates an interactive graphical network of concepts, enabling an easy navigation through this space. Selecting particular node shows its details and links to the related objects. The nodes and edges can be modified manually: the applet allows operations such as adding, editing and deleting components of the semantic space. The data changed in this way is marked as "manual", to distinguish the hand-crafted knowledge from learned or automatically generated knowledge. An example of such visual presentation of ontology-oriented semantic space is shown in Fig. 3.

b. Numerical semantic space representation

Neither the semantic wCRK structures, nor the database tuples, are useful for direct numerical calculations. However, the semantic space can be converted to the matrix of the Concept Description Vectors (CDV) – the numerical representation of relations for each concept [17]. The CDV vector components describe the strength of relations between particular keyword and the concept represented by the vector. Such very simple knowledge representation enables numerical processing of the information contained in the semantic space. Although it may not be sufficient for full parsing and understanding of texts it is useful for many applications. For example, semantic query system should understand what the user has in mind, and if this is not clear should ask minimum number of questions to gain additional knowledge. The 20 questions game serves as a good example for such question-answer applications [19], requiring an algorithm for guessing concepts that the user thinks about. In the simplest case the system is asking questions and the user answers 'yes' or 'no'. Using matrix representation of concept space where rows represent *M* objects (*o*) and columns represent *N* features (*c*) the best semantic space property maximizes information gain:

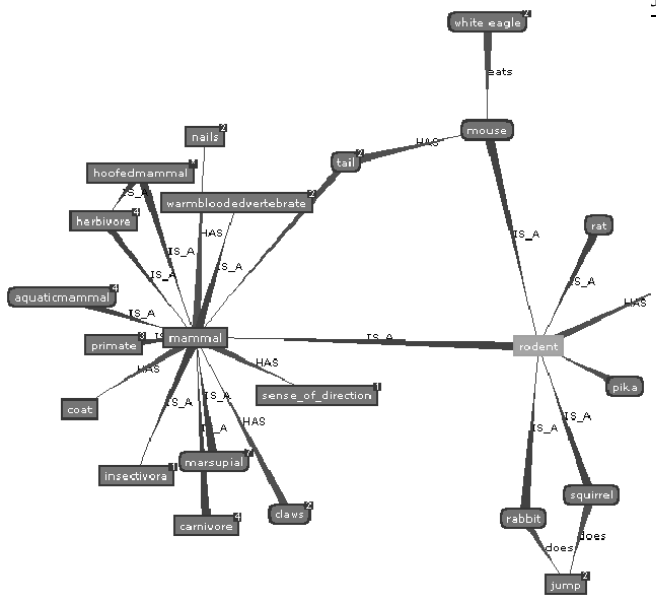

Fig. 3. Visualization of concept properties and relations.

$$IG(c_m) = -\sum_{i=0}^{N} p(o_i)\log p(o_i)$$

where $p(o_i) = \mathrm{abs}(w_{mi})/N$

Here w_{mi} is the strength of relation between object o an its feature c. If the weight is positive the keyword is relevant to the concept, if it is negative it is known that the concept does not have the particular property described by the keyword, and if the weight is 0 then the keyword is not applicable to the concept at all. The status may also be undetermined, the NULL relation weight codes for the lack of data.

In each step of the game feature that will maximize the information is calculated from the subspace of the most probable concepts. The relevant subspace is built from the initial query and the previous answers using the formula:

$$O(A) = \| CDV, ANS \| = \min$$

where the distance $\|\cdot\|$ between all concept description vectors (*CDV*) describing concepts, and the vector representing user answers (*ANS*), is minimal.

The initial semantic space is built with only "positive" relations, describing the concept using its properties. Thus the lack of relation with some feature is interpreted in negative sense, assuming that it implies that the concept has no such feature. Due to the knowledge incompleteness this assumption is frequently false and the knowledge in SM needs to be corrected in the learning process, as described below.

c. Natural dialog interface

The data stored as wCRK can be used to formulate simple sentences. The sentence generator (semSentenceGenerator) module in semAPI layer enables creation of simple queries based on selected wCRK. They serve as input data for the Humanized Interface (HIT) architecture incorporating three modules: Haptek [22] talking head, text to speech synthesis, and speech recognition. This interface is used in the human – machine dialog for data acquisition (Fig. 4 below).

The items imported into semantic memory from the WordNet lexicon are not directly useful in query precisiation or in word games. WordNet descriptions may contain many specialized terms (for example biological taxonomy terms) that are not known to most users, while a lot of knowledge that is obvious to humans is not explicitly mentioned. To verify and complete this data a version of the 20 questions game is used, and an interactive information exchange initiated with the users. Questions about particular semantic memory assertions (stored in CDVs) are formulated using specific dialog scenarios and answers used to improve representation of concepts in the semantic space [18]. Learning is realized through modification of weight values, estimating the strength of relations between concepts and features, and the certainty of such knowledge. If the assertion is true or false the strength is incremented or weakened, for confirmed relations certainty weights are increased.

Large weight between particular feature and concept means that the answer should be useful with high certainty, as the knowledge is about something widely known. Two active dialog scenarios have been implemented:

- Concepts acquisition: this is run when the SM-based system fails to guess the precise concept in the 20-question game. Using scenario: "I give up. Tell me what did you think of?" system can acquire a new concept. Representation of this concept in form of wCRK vector is formed using answers obtained during the game. For exiting objects SM system can correct strengths of relations of the concept with features that appeared during the game.

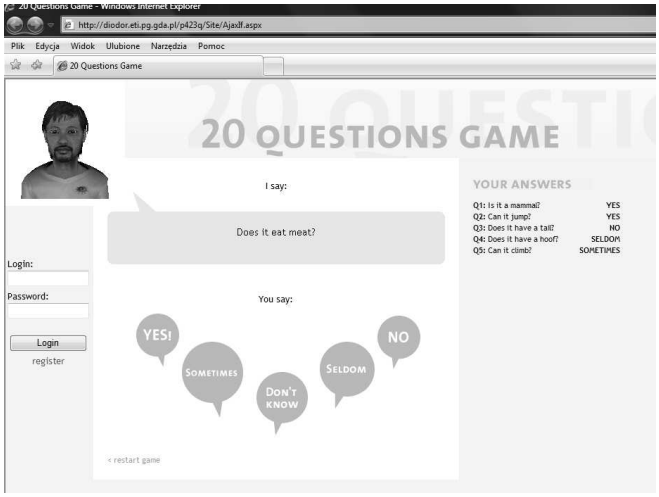

Fig 4. An example of the avatar based web interface available under IE at http://diodor.eti.pg.gda.pl

- Acquisition of new features: the second scenario "Tell me what is characteristic for <concept> " is used to separate two concepts that have very similar CDV representations.

This dialog is run when the SM system fails to discern some concepts during the game due to the lack of knowledge. Using it iteratively for similar concepts new descriptive features are introduced to the system.

These two simple scenarios allow to collect and purify SM knowledge. The learning process based on user answers obtained during games bootstraps on the existing knowledge and is as an alternative for handcrafting ontologies. Each of the finished games (failed or succeeded) causes data actualization – correction of weights describing relations.

IV. Experiments and results

The information retrieval system that asks relevant questions should help to precisiate the query to the point where only relevant documents will be retrieved or correct identification of object will be done. To illustrate how this approach may help to discover knowledge a simple experiment in medical domain has been performed. The DSM IV (Diagnostic and Statistical Manual of Mental Disorders) [23] contains 6 decision trees that allow for diagnosis in specialized psychiatric domains. It lists different categories of mental disorders and the criteria for diagnosing them. A search for information about particular mental problem based only on symptoms will be made.

DSM data has been used to construct the semantic space for relevant medical concepts. The diagnostic tree allows to compare our query-based algorithm against the human designed diagnostic process. Fig 5 presents the average number of queries made by the algorithm in each of the 6 subdecision tress, and in the aggregated data, compared to the sequential tests of the symptoms (questions about various aspects of the disease) used by the DSM manual. The average number of tests (questions) used in the standard diagnostic process (first bar), and used by the algorithm (middle bar) shows that the questions asked to increase information gain are nearly 45% better then the sequential tests in most decision trees, and about 20% better for the whole decision tree.

The query-based algorithm has also one more interesting feature. Increasing the margin that defines O(A) subspace allows to deal with mistakes in the answers. The extended margin (greater then minimum) helps to correct discrepancy between test results and knowledge stored in semantic memory. If some answers are false the system needs to ask more questions, but the standard decision tree algorithm will completely miss the correct decision. The third bars in Fig. 5 show the average number of questions asked by the algorithm when the margin = 1 is used, implying the ability to work with one mistake. The number of questions in all decision trees is now larger than in the previous two cases because extended margin allows for bigger incompatibility between CDV and ANS vectors, and therefore more tests should be done to separate the most probable concept from all other concepts.

Although the algorithm will work also in case of larger number of mistakes even more questions will be needed to correct these mistakes during decision process. In real life also the expert may notice that diagnosis can't be reached, the answers have not been consistent and therefore one should return and check which questions have been mistakenly answered.

Verification and improvement of the quality of SM knowledge using the 20 questions game has also been tested on the semantic space created for animal kingdom domain. The demonstration of the algorithm in this domain is available at the **diodor.eti.pg.gda.pl** website. Specialized biomedical or technical domains may be more impressive, but animal domain is better for tests, because it is quite large while still easy to understand, as most people share similar knowledge related to animals. These results may be used in a search engine that will help to find precise animal the user has in mind, for example watching a picture of unknown animal. There may be many applications of this sort, for example helping to recognize plant names, a process that requires a lot of effort if traditional search engines are used.

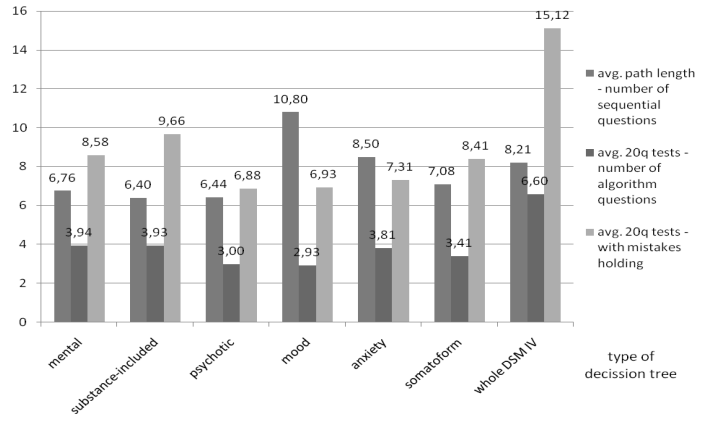

Fig 5. Average number of queries in different DSM IV decision tress using 3 different approaches.

V. Future developments

The semantic web approach requires a lot of manual effort to build RDF-based ontologies, and a lot more effort if DAML+OIL or other sophisticated description languages are used. In effect not so many ontologies have been build in specialized domains. Search engines are based on links and statistical relations, with a few engines dedicated to gene ontologies or other narrow subjects, that actually use ontologies for their background knowledge. As a result information retrieval is frequently quite difficult. Humans are far better in answering questions rather than providing good descriptions of whatever they search for.

The importance of ontologies in information systems will certainly grow. Lack of well defined common sense ontologies or semantic memories that could easily acquire new knowledge, and lack of good algorithms to build them, is a

major obstacle in improving the natural language interfaces. This is clearly seen for hand-crafted lexical resources such as the WordNet, a collective long-term effort of many people that still contains a lot of accidental information, sometimes very specialized knowledge that only a few people would understand, while information that is obvious to humans is missing. For example, all people know how a 'horse' looks like, so there is no need to mention explicitly in dictionaries or encyclopedias that horses have ears, tail or hoofs, but 'canon' or 'shank' body parts are described, although few people know such concepts.

The slow progress of practical applications that use natural language processing beyond a simple search has brought some experts to the idea that real understanding of concepts cannot be achieved on the purely linguistic level, and that grounding of the semantics in real internal representation of visual, auditory and tactile experiences of a robot is necessary. This may be too pessimistic, as still even the basic things, such as creation of common sense ontologies, have not been done. Our approach is to create, using machine learning techniques, an initial representation of semantic memory based on concepts described by the wCRK triples and weights, and for some applications to simplify this representation further to the CDV form. Such knowledge items may contain wrong information about concepts. The active learning algorithm may be used for verification of common-sense ontologies. The concept visualization tools (Fig. 3) may also be used to correct knowledge directly. This approach to build knowledge bases should be more efficient (although not without some limitations) than the declarative approaches used in such large-scale systems as CyC [24].

To gain quickly a lot of common-sense knowledge that should be useful in information retrieval a large-scale collaborative project is planned, utilizing search competitions for information about various concepts, as well as learning from word games played by many users, and active dialogs with chatterbots that may ask directly questions about objects mentioned in the dialogue. Most chatterbots are still based on identification of keywords and template-matching algorithms, the same technique as has already been used by the Eliza program of Weizenbaum [25].

The approach to information retrieval based on asking minimum number of questions to precisiate the concept will obviously fail if the search is concerned with an answer to a complex question, for example the reason for an error of computer system in some specific situation. Understanding the problem may then require complex parsing of the question that only a real expert may be capable of. However, in most common situations asking minimum number of relevant questions will lead to the desired information faster than text engine searches or clustering techniques.

References

[1] Berners-Lee T, Hendler J. and Lassila O, The Semantic Web. Scientific American, May 17, 2001.

[2] Davies J, Semantic Web Technologies: Trends and Research in Ontology-based Systems. J. Wiley 2006.

[3] Handschuh S, and Staab S. (eds.). Annotation for the Semantic Web. IOS Press, 2003.

[4] RDF description, see: http://www.w3.org/RDF/

[5] Usui S, Visiome: neuroinformatics research in vision project. Neural Networks 16(9), 2003 , pp. 1293-1300

[6] Schatz, B.R., Information Retrieval in Digital Libraries: Bringing Search to the Net, *Science* 275(5298), 327-334

[7] Schatz B, The Interspace: Concept Navigation across Distributed Communities, IEEE Computer, 35(1): 54-62, 2002.

[8] Bennett N, He Q, Powell K. and Schatz B. Extracting Noun Phrases for All of MEDLINE, Proc. of American Medical Informatics Assoc. Annual Conf 1999, Washington, DC, pp. 671-675.

[9] Lagus K, Kaski S, and Kohonen T. Mining massive document collections by the WEBSOM method. Information Sciences, 163, 135-156, 2004.

[10] Zadeh L.A, Precisiated natural language (PNL), AI *Magazine* 25, 74–91, 2004.

[11] Tulving E, Episodic and Semantic Memory; in: Tulving, E, Donaldson W (Eds): Organization of Memory. New York 1972.

[12] Smith, E.E., Shoben, E.J., and Rips, L.J. Structure and process in semantic memory: A featural model for semantic decisions. Psychological Review, 81, 214 241, 1974.

[13] Sowa, J.F. ed. Principles of Semantic Networks: Explorations in the Representation of Knowledge. Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[14] Collins A.M. and Quillian M.R, Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior 8, 240-7, 1969.

[15] Collins A.M. and Loftus E.F, A spreading-activation theory of semantic processing. Psychological Reviews 82, 407-28, 1975.

[16] Wordnet, see: http://wordnet.princeton.edu

[17] Szymanski J, Duch W, Knowledge representation and acquisition for large-scale semantic memory. World Congress on Computational Intelligence (WCCI'08), Hong Kong, 1-6 June 2008, IEEE Press (in print)

[18] Szymanski J, Duch W, Semantic Memory Architecture for Knowledge Acquisition and Management. 6th International Conference on Information and Management Sciences (IMS2007), July 1-6, 2007, California Polytechnic State University, CA, pp. 342-348.

[19] Szymanski J, Sarnatowicz T, Duch W, Towards Avatars with Artificial Minds: Role of Semantic Memory. Journal of Ubiquitous Computing and Intelligence, American Scientific Publishers 2, 1-11, 2008.

[20] Duch W, Matykiewicz P, and Pestian J, Neurolinguistic Approach to Natural Language Processing with Applications to Medical Text Analysis. *Neural Networks* (in print, 2008)

[21] Touchgraph, see: http://www.touchgraph.com/

[22] Haptek avatars, see: http://www.haptek.com

[23] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, 1980 APS

[24] Lenat D.B, CYC: A Large-Scale Investment in Knowledge Infrastructure. Comm. of the ACM 38, 33-38, 1995.

[25] Weizenbaum J, Computer Power and Human Reason: From Judgment to Calculation. W. H. Freeman & Co. New York, USA 1976.

Julian Szymański received the BEng and MSc degrees from Gdańsk University of Technology in computer science and from the Nicolaus Copernicus University in philosophy; he works currently as the teaching assistant at Gdańsk University of Technology.

Włodzisław Duch received the MSc, PhD and DSc degree from the Nicolaus Copernicus University, and is the Head of Department of Informatics at this university; recently (2003-2007) he has worked as a Visiting Professor at Nanyang Technological University, Singapore. Currently he serves as the President of the European Neural Networks Society (www.e-nns.org).

For more information Google: W. Duch.