ConceptWorld: Uncovering the Pragmatic Depth of Semantic Concepts Using the WWW

Michael Dittenbach[1], Helmut Berger[1], Dieter Merkl[2], and Florian Eckkrammer[1]

[1] E-Commerce Competence Center – EC3
Donau-City-Strasse 1, A–1220 Wien
{michael.dittenbach, helmut.berger, florian.eckkrammer}@ec3.at
[2] Institut für rechnergestützte Automation, Technische Universität Wien
Karlsplatz 13/183, A–1040 Wien
dieter.merkl@inso.tuwien.ac.at

Abstract. The task of researching information on a particular topic using the Web is mainly accomplished by relying on keyword-based search engines. Although this approach provides a good starting point, it remains a tedious task to collect additional information that puts this topic in greater context. In this paper we present ConceptWorld, an instrument to automatically discover various facets of a topic of interest by extracting concepts from Web documents. The result materializes as a network of semantic concepts with their various contextual interrelations and provides a holistic view on the topic of interest.

1 Introduction

It is commonplace to view the World Wide Web as a symbolic system. Its symbols are Web page content and markups. As structural aspects of the Web have become better understood, increasing attention is drawn towards semantics. To bring the Semantic Web to life, however, efficient ways to access and extract semantic concepts from Web documents are needed. A *semantic concept* is a natural language fragment that is semi-structured according to an ontology of allowable syntactic patterns, e.g. phrases such as "Multiple Sclerosis", "Microsoft Windows" or "Franz Ferdinand". When researching for information about a certain topic or concept, it is vital to capture its various context-sensitive aspects of meaning in order to obtain a holistic view [1]. This materializes as a network of semantic concepts and their various contextual interrelations.

In principle, Web page annotations could facilitate the identification of semantic concepts and their relations to other concepts. However, annotations – especially those exceeding a certain level of quality – are rare and will probably never be rich or detailed enough to cover all the context-sensitive aspects of meaning of the semantic concept. Manual annotation is impractical and unscalable, and automatic annotation tools remain largely undeveloped [2].

To gain independence from manually created annotations, we have developed ConceptWorld, an instrument to identify semantic concepts from natural

language Web documents. ConceptWorld facilitates the discovery of a concept's context-sensitive aspects of meaning and its various contextual interrelations. Our basic hypothesis is that the Web already contains information in natural language documents about virtually every thinkable topic. We may thus see the Web as a universal encyclopedia of admittedly highly divergent quality [3]. Thus, with ConceptWorld it is possible to obtain a holistic view on arbitrary topics. As an example consider the acronym *MS* which is used in a variety of contexts, such as *Microsoft*, *Mississippi* or *Master's degree*. *MS* also refers to *Multiple Sclerosis* in medicine and we will use this long form as an example throughout the paper. It may be described in a sentence such as "Multiple Sclerosis is a disease of the central nervous system", which can be decomposed into a syntactic 3-tuple comprising the noun phrase (subject) "Multiple Sclerosis", the verb "is" and the noun phrase (object) "a disease of the central nervous system". We denominate both noun phrases as concepts. In order to discover the various facets of this particular topic, ConceptWorld extracts objects from a multitude of Web documents, associates them with *Multiple Sclerosis* and, thus, provides an initial description of the topic. Subsequently, the predicate and each identified object are used to obtain associated Web documents containing sentences such as "Alzheimer is a disease of the central nervous system". This time, though, the subject "Alzheimer" is extracted and is associated with the object "a disease of the central nervous system". As a result, ConceptWorld provides a network of semantic concepts centered around *Multiple Sclerosis*.

In a nutshell, ConceptWorld is a research instrument to discover the pragmatic depth of a semantic concept rather than a search engine. Thus, ConceptWorld transcends traditional search engines that merely provide pointers to documents containing the textual representation of the semantic concept while disregarding its context-sensitive aspects of meaning. Pragmatically speaking, it is a non-trivial task to start your research with *Multiple Sclerosis* and to discover its relationship to *Alzheimer's Disease* when relying on contemporary search engines. The discovery and representation of relationships between semantic concepts, say, *Multiple Sclerosis* and *Alzheimer's Disease*, is a designated feature of ConceptWorld.

The remainder of this paper is structured as follows. In Section 2 we provide an overview on related approaches. The idea underlying ConceptWorld is described in Section 3 along with experimental results in Section 4. A discussion of the current approach and an outlook on future work is presented in Section 5. A conclusion is given in Section 6.

2 State of the Art

Currently, the task of researching information on a particular topic using the Web is mainly accomplished by relying on keyword-based search engines. Google (`http://www.google.com`), for instance, returns a ranked list of about 22 million Web pages when querying for the phrase "Multiple Sclerosis". Although the top-ranked pages provide a good starting point for the research task and

contain detailed information about Multiple Sclerosis, it remains a tedious task to assemble additional information that puts this topic in greater context. So it is difficult to discover the context-sensitive aspects of the meaning of Multiple Sclerosis and its various contextual interrelationships with other concepts such as Alzheimer's Disease. Google's *Similar Pages* feature provides a set of Web pages about somehow related diseases, but, however, does neither reveal the type of relationship nor is their interrelationship presented transparently. The interpretation is left to the searcher.

Mooter (`http://www.mooter.com`) goes a step further and combines keyword-based search with clustering of results based on the pages' contents. In addition, the clusters are automatically labeled to hint at the common grounds of the single clusters. Although this is an improvement over simple lists regarding result presentation, the pages found are still limited to those exactly containing the keywords or key phrases. For "Multiple Sclerosis", for instance, the most meaningful cluster labels on the first result page are "disease" and "treatment", each of which contains the set of pages where Multiple Sclerosis is described in the context of being a disease and in the context of its treatments, respectively. The pages of a particular cluster have in common that the user's original query and the respective cluster label are present.

So far we have described contemporary Web search engines that more or less rely on word occurrences and link analysis between Web pages [4]. We have to refer to the Semantic Web as an orthogonal approach, which is envisioned to create a universal medium for information exchange by semantically annotated documents with computer-processable meanings [5]. Currently, methods and technologies are researched and developed to provide the technological framework facilitating document annotation. The Resource Description Framework (RDF), for instance, provides a language for modeling semi-structured metadata and enables knowledge-management applications. Additionally, a lightweight schema language, RDF Schema (RDFS), offers basic structures such as classes and properties. However, since RDF has expanded to accommodate the Semantic Web, the limitations of RDFS have become evident which led to the development of a more expressive schema language, the DARPA Agent Markup Language (DAML). The Ontology Inference Layer (OIL), intended to provide even more sophisticated classification, is a Web-based representation and inference layer for ontologies which combines the widely used modeling primitives from frame-based languages with the formal semantics and reasoning services provided by description logics. This materialized in DAML+OIL, a language for expressing sophisticated classifications and properties of arbitrary resources. For a concise overview of the Semantic Web and its languages we refer to [6].

Despite all efforts in developing technologies to support authoring of Semantic Web documents, the lack of semantically annotated documents is evident. To paraphrase McCool [7], the current Semantic Web remains a parallel universe in the shadow of the World Wide Web. Swoogle (`http://swoogle.umbc.edu`), is one among the few retrieval systems for the Semantic Web [8]. In a nutshell, Swoogle extracts metadata for each crawled document, computes relations be-

tween documents and provides access to these documents via a search interface. Currently, the retrieval system's index comprises 1.5 million Semantic Web documents; a rather small number compared to approximately 8 billion Web pages indexed by Google as of March 2005. In the light of these figures, it is reasonable to conclude that manual annotation is impractical and unscalable. Thus, tools that take advantage of semantic information implicitly contained in natural language Web documents are needed.

3 ConceptWorld

ConceptWorld is an instrument to automatically identify semantic concepts that relate to a particular source concept. Natural language Web documents are the basis for the discovery of the context-sensitive aspects of meaning of these concepts and their various contextual interrelations. The result obtained with this instrument is a network of semantic concepts. The internal representation of the ConceptWorld network can be considered as a weighted, directed graph. Formally, let $CW = \langle C, R \rangle$ be a pair, where C is a set of concepts (vertices) and $R = \{(c_i, c_j) | c_i, c_j \in C \wedge c_i \neq c_j, w \in \Re\}$ is a set of weighted relations (edges) between concepts.

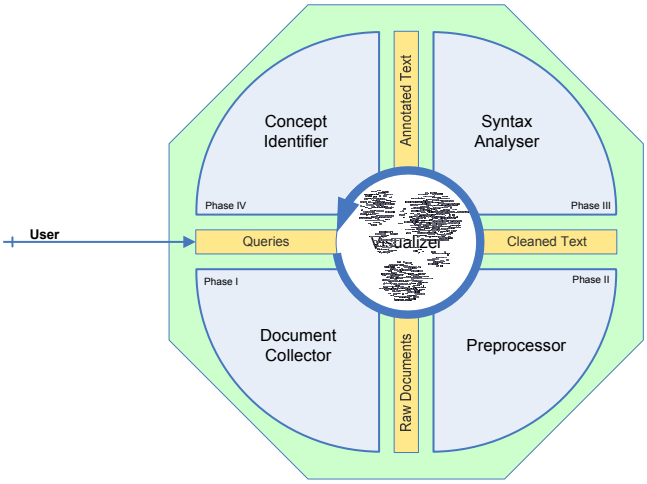

Fig. 1. The ConceptWorld process loop.

The idea underlying ConceptWorld is depicted in Figure 1. This idea materializes in a process loop consisting of four phases which are applied in two iterations. The first phase, i.e. document collection, relies on several well-known search engines including but not limited to Google, Yahoo and AltaVista. A search engine query is composed by concatenating the source concept $s \in C$, the relation expressed in terms of the verb "is" and a conjunction of the determiners

"a", "an" and "the". For Google, this leads to a query in the form of "s is (a OR an OR the)" including the double quotes to force the search engine to retrieve only those documents containing the exact sentence fragment. The exact syntax for linking the determiners with Boolean operators depends on the respective search engine, and hence, may vary slightly. The purpose of adding the determiners to the query is to retrieve documents that contain sentences where the particular concept – usually represented by a noun phrase in linguistic terms – is again described in terms of a noun phrase rather than an adjective. In other words, we want to find statements describing *what* rather than *how* things are. ConceptWorld is not limited to a specific relation type such as the *is-a* relation. It is, of course, possible to use different patterns expressing other relation types to create networks of concepts explaining, for example, what things *mean*, who *is a relative of* whom, or which artist *was inspired by* another.

In the next step, the query is submitted to the search engines. We generate the union of the n top-ranked URLs returned by each search engine to obtain a set of links to pages containing the query. Note that advertised or sponsored links are discarded. Then, the Web resources identified by the URLs are downloaded in parallel and stored locally. Currently, these include plain text, HTML pages, PDF and RTF documents. In order to circumvent the problem of latency caused by congested network connections, the download process is terminated after a particular timeout.

In the second phase the documents are preprocessed according to their MIME types. We distinguish three types of preprocessors. Plain text documents are passed through without any modification. In case of PDF and RTF documents, we rely on external conversion tools such as `pdftotext` and `rtf2text`. The tags contained in HTML documents are removed with a number of exceptions. Headings, list elements or paragraph separators, for instance, are processed such that each opening tag is removed but the closing tag is replaced by a period. This is done to provide support for the natural language processing steps in the subsequent analysis phase. The result of Phase II is a plain text representation of each document.

The third phase comprises three natural language processing tasks. First, each plain text document is split into sentences. Second, only those sentences containing the query are selected for further processing. Third, a part-of-speech tagger (`http://search.cpan.org/~acoburn/Lingua-EN-Tagger/`) including a chunking algorithm is used to syntactically annotate and group the components of the sentences. The result is a set of syntactically annotated sentences.

In Phase IV, the first noun phrase *after* the verb is selected. Our heuristic assumption is that this noun phrase represents a concept c describing the source concept s with respect to the underlying *is-a* relation. A new concept c is added to the graph, if c is not element of C. Analogously, a new relation $r = \{(s, c), w\}$ between the source concept s and c is added, if r is not element of R. In case of a new relation its weight w is set to 1, otherwise the weight of the existing relation is increased by 1 if the corresponding Web document, i.e. URL, has not yet contributed to this particular relation.

The second iteration of the process loop is applied to each concept identified in the first iteration, i.e. $c \in \{C \setminus s\}$. The process is basically the same as with the source concept, however, the following two differences apply.

1. A query is assembled by concatenating the relation, i.e. verb and determiners, with concept c as object, resulting in the query "is (a OR an OR the) c".
2. In Phase IV, the first noun phrase *before* the verb is selected as opposed to the first noun phrase *after* the verb. This time, our heuristic assumption is that this noun phrase represents a concept t that is described by the underlying *is-a* relation to concept c, which was identified during the first iteration.

Phases I to III, i.e. document collection, preprocessing and syntactical analysis, are carried out in exactly the same way as during the first iteration. Again, the new concept t is added to the graph, if $t \notin C$. Then, a relation $r = \{(t, c), w\}$ is added, if $r \notin R$. In case of a new relation the weight w is initialized with 1, otherwise it is increased by 1. The result of this process is a graph structure centered around the source concept s with connections to concepts describing s in terms of *what* it *is*, which are, in turn, connected to other related concepts. The XML representation of the graph CW is passed on to the Visualizer, which, in turn, graphically represents the concepts in terms of a semantic network.

4 Results

We illustrate the power of ConceptWorld by means of *Multiple Sclerosis* as source concept for a hypothetical research task. Note that we used the 25 top-ranked URLs returned by each search engine to obtain a set of links to pages containing "Multiple Sclerosis". The resulting ConceptWorld graph contains 1,512 concepts connected via 1,768 edges. For the sake of clarity, we refrain from depicting the complete graph and focus on two distinct types of concepts selected according to their degree. In Figure 2, only those concepts are shown that i) are direct neighbors of the source concept and ii) have a degree of one, and are thus not connected to any other concept. These very specific concepts provide a fairly concise picture of the disease and it can be expected that they occur only in combination with the phrase "Multiple Sclerosis". A closer look confirms this hypothesis as rather long fragments containing highly descriptive definitions, such as "organ specific autoimmune disease orchestrated by autoreactive t cells" or "central nervous system inflammatory demyelinating disease", are obtained.

Figure 3 shows the subgraph containing only those neighbors of the source concept with a degree of equal or more than 8, representing the antipodal type of concepts. These hub concepts exhibit a rather high level of connectivity and provide a generalized view on the source concept. Examples of such hub concepts are *disease*, *disorder*, *autoimmune disease* and the like.

In general, these hub concepts relate to a variety of medical terms, see Table 1 for an illustration. A closer look on the concept "disease", however, reveals that its usage is not exclusively limited to medicine. ConceptWorld collected a Web

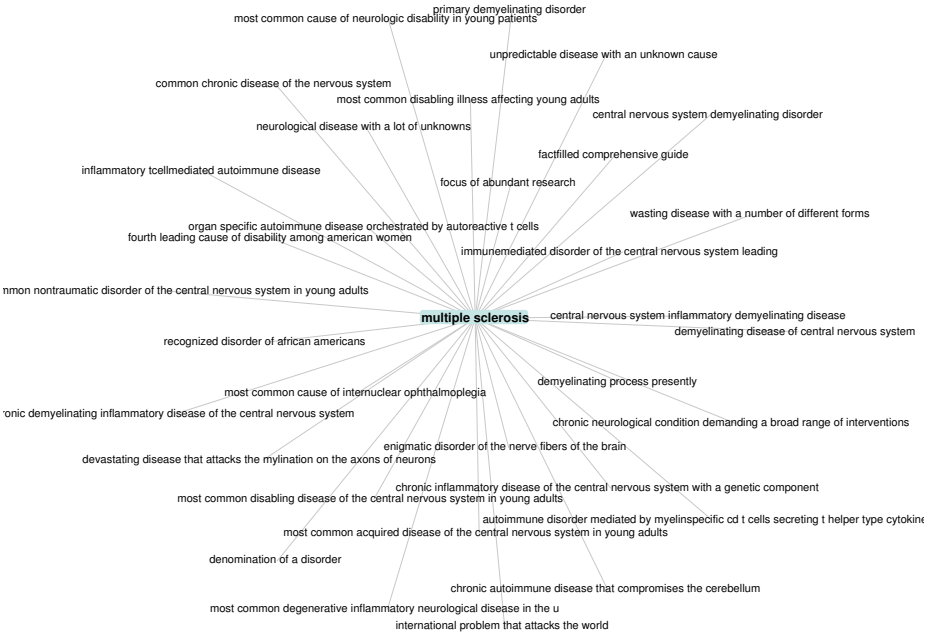

Fig. 2. ConceptWorld graph for "Multiple Sclerosis"; degree = 1.

document (`http://pandagon.net/2005/07/13/`) containing the following sentence fragment: "It's standard issue fascist thinking that liberalism is an disease [...]". Other sentences with similar patterns were discovered forming a cluster of concepts related to "disease" including political themes, such as "conservatism", "liberalism", "capitalism", "communism", "antisemitism", or "politics". Moreover, the range of concepts regarded as related to "disease" are captured with the following samples: "truth", "boredom", "terrorism", "poverty", "beauty", "loss of spiritual equilibrium, "homosexuality", "war", "bigotry" or "aging". Note that we refrain from showing the subgraph of the concept "disease" since the number of 227 neighbors would render the illustration unreadable.

Table 1. Sample concepts related to *Multiple Sclerosis* via different hub concepts.

hub concept	**medical concepts**
disease	zoonosis, depression, diabetes, parkinson, alzheimer, ...
autoimmune disease	grave's disease, crohn's disease, alzheimer, systemic lupus erythematosus, ...
disease of the central nervous system	reflex sympathetic dystrophy, complex regional pain syndrome (rsd crps) chronic wasting disease (cwd), alzheimer, parkinson, poliomyelitis, ...
disorder	bronchopulmonary dysplasia (bpd), paranoia, epilepsy, dyslexia, sleep apnea, ...
tragic illness	parkinson's disease, hiv, schizophrenia, alzheimer's disease (ad), alcoholism, ...

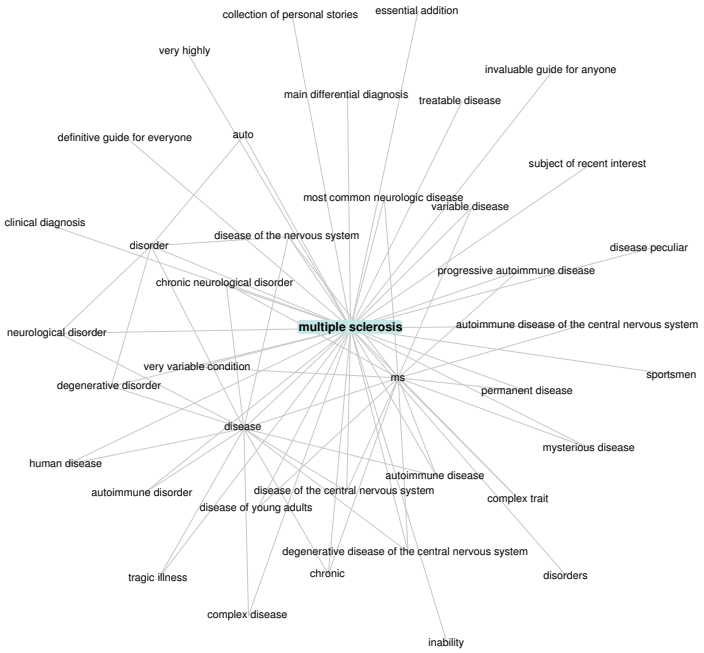

Fig. 3. Hub concepts with degree >= 8 for "Multiple Sclerosis".

In a similar way, the graph of the hub concept "disorder" identifies 83 concepts related to abnormal physical or mental conditions, see Figure 4. This includes a set of syndromes such as "chronic fatigue syndrome", "restless legs syndrome", "fetal alcohol syndrome", neurological disorders such as "narcolepsy", "pervasive developmental disorder (pdd)" or "epilepsy". However, some net citizens also perceive "order" or "windows devices" to be a type of "disorder". Mathematical concepts such as "symmetry order" or "one order type" are regarded to be a "disorder" as well. These examples are of course illustrative rather than exhaustive.

5 Discussion

An important point to note is that the relations extracted with ConceptWorld do not necessarily reflect the *absolute truth* about a particular topic. ConceptWorld's holistic view represents what people, i.e. the publishers of Web pages, write about their perceptions of the world they live in. To cut short a philosophical discussion, a ConceptWorld graph created for a particular topic contains relations that are explicitly expressed in terms of written, natural language texts. It is not our aim to verify the correctness of the statements extracted from the retrieved documents. A quite obvious example from the information technology domain is shown by the results of a ConceptWorld process for "Microsoft

Fig. 4. Hub concept "disorder".

Windows" with the "is-a" relation. The ConceptWorld graph includes the *true* relation to "operating system", which further relates to all types of operating systems such as "Linux", "MacOS", "Plan9" and the like. Another sample relation would be "trademark of Microsoft" that, in turn, has relations to "Windows Media" or "Excel". On the other hand, there exist relations to "force of nature", "complicated system" or "target of attack for viruses and Malware", which rather show what people *think* about Microsoft Windows. This, again, underlines the genuine property of ConceptWorld of finding relations in different contexts distinguishing our approach from the search engines mentioned earlier, which limit their results to pages containing the exact query terms or phrases. With the second iteration of the ConceptWorld process, the scope of the information space limited by the keywords describing the source concept is transcended.

In the context of "Multiple Sclerosis" this results in the inclusion of a broad spectrum of autoimmune diseases such as "Crohn's Disease", "Lupus", "Berger's Disease" or "HIV", see Figure 5. Interestingly, the Human Immunodeficiency Virus (HIV), a retrovirus that is the cause of the disease known as AIDS, is linked to the concept "autoimmune disease". This connection is the result of the wide-spread misconception of the relation between HIV and AIDS, i.e. AIDS is the autoimmune disease not HIV. Strictly speaking, this reflects people's mis-

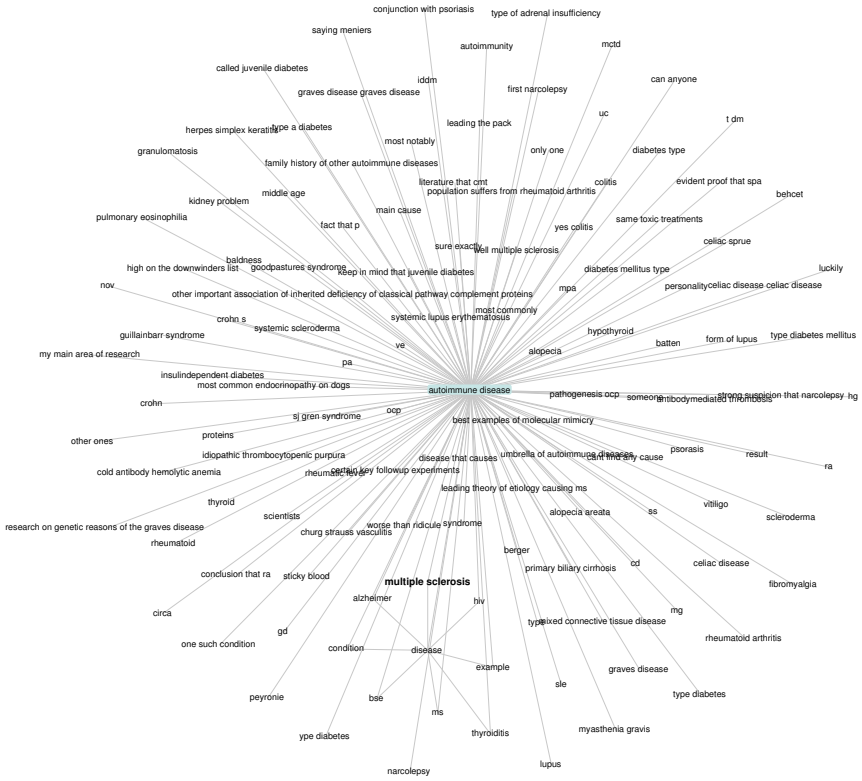

Fig. 5. Hub concept "autoimmune disease" with 125 neighboring concepts.

taken perception of HIV. The thematic coverage of the graph is not limited to autoimmune diseases of humans, it also includes diseases among animals. As an example consider "Bovine Spongiform Encephalopathy (BSE)" or mad cow disease which is a chronic, degenerative disorder affecting the central nervous system of cattle.

Another observation is the wide-spread use of acronyms and abbreviations in medicine. The resolution of concepts such as "ra" or "mg" into their corresponding long form, i.e. rheumatoid arthritis and myasthenia gravis respectively, will increase the expressiveness of the graph.

However, the graph also includes some peculiar terms such as "leading the pack". This particular concept is extracted from the sentence "Leading the pack is the autoimmune disease Sjogren's syndrome, which impairs lacrimal gland function and the formation of watery tears." published on a particular Web page (`http://www.diabetic-help.com/mmdryeye.txt`). The heuristic assumption followed during the second iteration of the ConceptWorld process identified "leading the pack" as being the noun phrase of interest. This might be addressed with deeper natural language analysis of the sentence.

Another potential improvement is exemplified by means of the concept "graves disease graves disease". Actually, the first occurrence refers to the heading of the article and the latter one identifies the beginning of the paragraph. Since the closing tag of the heading was missing, it was not replaced by a period during the analysis phase. So, the sentence splitter failed to correctly separate these two noun phrases due to the error-proneness of the HTML source code of the Web page. The identification of noun phrase replications and their subsequent merging will positively effect the number of concepts of the graph and help to render it more comprehensible.

The temporal dimension plays another important role when using an information source as ever-changing as the Web. The emergence of particular concepts described by words and names that are already occupied, can lead to disappearance or at least reduced visibility of older concepts from a search engine's perspective. Since about 2001, for example, it is rather difficult to find information about the Austrian Archduke Franz Ferdinand, whose assassination in Sarajevo has been an important event en route to WWI, since it is scarcely present among the top-ranked search engine results when using "Franz Ferdinand" as query. This is due to the increased popularity of the Scottish rock band of the same name. Consequently, ConceptWorld is also effected by the *zeitgeist* driving the creation of a significant amount of the content published on the World Wide Web. One way to address this complexity would be to generate a mixture of high-, middle- and low-ranked documents instead of using only the top-ranked documents during the document collection phase of ConceptWorld.

Besides these content-dependent issues, a number of technology-related approaches to improve the quality of the ConceptWorld process are worth mentioning. To increase the number of sentences containing a matching statement expressing the relation between two concepts, multi-sentential anaphora resolution can be introduced in the third phrase of the ConceptWorld process. Prior to splitting a document into single sentences, paragraphs will be analyzed for substituting personal pronouns with the concept they refer to. Consider the following snippet from a Web page (`http://www.mira.ca/contenta/s1-2a.html`): "[...] Multiple sclerosis affects 2 times more women than men. It is the central nervous system disease the most extended for the young [...]". Anaphora resolution would detect that *It* in the latter sentence refers to *Multiple Sclerosis*. Hence, the Concept Identifier extracts the concept described by the phrase "central nervous system disease".

Currently, the ConceptWorld graph for "Multiple Sclerosis" includes relations to "incurable degenerative disease", "chronic and often disabling disease", "complex disease" or "unpredictable disease". Noun phrases may also contain conjunctions, adverbs or adjectives. To this end, a heuristic for merging such phrases based on the part-of-speech tags attributed to the words can be used to improve the manageability and clarity of the graph. To tackle the syntactic and semantic challenges related to natural language processing, we will integrate tools such as WordNet [9] or GATE [10]. Additionally, we will investigate the ex-

tent to which commonsense reasoning, as provided for example by MontyLingua [11, 12], can further improve the quality of the current ConceptWorld approach.

6 Conclusion

In this paper we have described ConceptWorld which facilitates the discovery of the various facets of a topic of interest. ConceptWorld extracts concepts from Web documents determined with search engines. The result is a semantic network of concepts that provides a holistic view representing people's perceptions of the world they live in. In a nutshell, ConceptWorld is a research instrument to discover the context-sensitive aspects of meaning of a semantic concept.

We have demonstrated the advantage of ConceptWorld by means of the topic "Multiple Sclerosis". The thematic coverage ranged from definitions of the disease, over a number of related autoimmune diseases and disorders, even to political statements. This highlights the genuine property of ConceptWorld of finding relations in different contexts distinguishing our approach from contemporary search engines.

References

1. Marchionini, G.: Exploratory search: From finding to understanding. Communications of the ACM **49**(4) (2006) 41–46
2. Schoop, M., de Moor, A., Dietz, J.: The pragmatic web: A manifesto. Communications of the ACM **49**(5) (2006) 75–76
3. Cilibrasi, R., Vitanyi, P.: Automatic meaning discovery using google. Technical report, CWI, University of Amsterdam (2004)
4. Henzinger, M.: Hyperlink analysis for the web. IEEE Internet Computing **5**(1) (2001) 45–50
5. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **279**(5) (2001) 34–43
6. Fensel, D.: The semantic web and its languages. IEEE Intelligent Systems **15**(6) (2000) 67–73
7. McCool, R.: Rethinking the semantic web, part 1. IEEE Internet Computing **9**(6) (2005) 86–88
8. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: A search and metadata engine for the semantic web. In: Proceedings of the 13th ACM International Conference on Information and Knowledge Management, Washington, D.C., USA, ACM Press (2004) 652–659
9. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. The MIT Press (1998)
10. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. (2002)
11. Liu, H.: MontyLingua: An end-to-end natural language processor with common sense (2004) Available at: `http://web.media.mit.edu/~hugo/montylingua`.
12. Liu, H., Singh, P.: ConceptNet: A practical commonsense reasoning toolkit. BT Technology Journal **22**(4) (2004) 211–226