# Automated Concept Discovery from Web Resources

Michael Dittenbach, Helmut Berger
iSpaces Research Group
E-Commerce Competence Center – EC3
Donau-City-Strasse 1, A–1220 Wien, Austria
{michael.dittenbach,helmut.berger}@ec3.at

Dieter Merkl
Institut für rechnergestützte Automation
Technische Universität Wien
Karlsplatz 13/183, A–1040 Wien, Austria
dieter.merkl@inso.tuwien.ac.at

## Abstract

*The task of researching information on a particular topic using the Web is mainly accomplished by using 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[1], it is vital to capture its various context-sensitive aspects of meaning in order to obtain a holistic view [6]. This materializes as a network of semantic concepts and their various contextual interrelations. Web page annotations could facilitate the identification of semantic concepts and their relations to other concepts. However, annotations – especially those above a certain quality level – 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 [8].

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 highly divergent quality [2]. Thus, with ConceptWorld it is possible to obtain a holistic view on arbitrary topics.

Currently, the task of researching information on a particular topic using the Web is mainly accomplished by relying on keyword-based search engines. Google, 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. Search engines such as Mooter (`http://www.mooter.com`) or Kartoo (`http://www.kartoo.com`) go a step further and combine keyword-based search with clustering of results based on the pages' contents.

So far we have described contemporary Web search engines that more or less rely on word occurrences and link analysis between Web pages [5]. We have to refer

[1] Please note that we use the terms *topic* and *concept* interchangeably throughout this paper.

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 [1, 4]. 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 [3]. In a nutshell, Swoogle extracts metadata for each crawled document, computes relations between 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 Web documents are needed.

## 2 ConceptWorld

ConceptWorld is an instrument to 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.

The idea underlying ConceptWorld 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 possible to use different patterns expressing other relation types to create networks of concepts explaining, for example, what things *mean* 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. 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 converted to plain text according to their MIME types. 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 to increase the quality of the sentence splitting algorithm used in the subsequent phase.

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 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, two differences apply. Firstly, 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) $c$". Secondly, 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 $t$

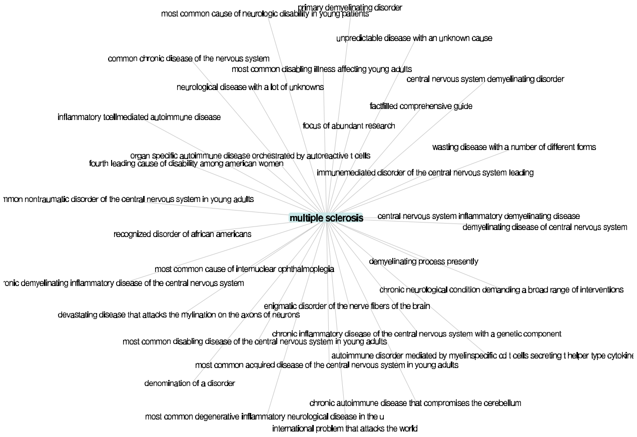
primary demyelinating disorder
most common cause of neurologic disability in young patients
unpredictable disease with an unknown cause
common chronic disease of the nervous system
most common disabling illness affecting young adults
central nervous system demyelinating disorder
neurological disease with a lot of unknowns
factfilled comprehensive guide
inflammatory tcellmediated autoimmune disease
focus of abundant research
wasting disease with a number of different forms
organ specific autoimmune disease orchestrated by autoreactive t cells
fourth leading cause of disability among american women
immunemediated disorder of the central nervous system leading
mmon nontraumatic disorder of the central nervous system in young adults
multiple sclerosis
central nervous system inflammatory demyelinating disease
demyelinating disease of central nervous system
recognized disorder of african americans
demyelinating process presently
most common cause of internuclear ophthalmoplegia
ronic demyelinating inflammatory disease of the central nervous system
chronic neurological condition demanding a broad range of interventions
enigmatic disorder of the nerve fibers of the brain
devastating disease that attacks the mylination on the axons of neurons
chronic inflammatory disease of the central nervous system with a genetic component
most common disabling disease of the central nervous system in young adults
autoimmune disorder mediated by myelinspecific cd t cells secreting t helper type cytokine
most common acquired disease of the central nervous system in young adults
denomination of a disorder
chronic autoimmune disease that compromises the cerebellum
most common degenerative inflammatory neurological disease in the u
international problem that attacks the world

**Figure 1. ConceptWorld graph for "Multiple Sclerosis"; degree = 1.**

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.

## 3 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. We focus on two distinct types of concepts selected according to their degree. In Figure 1, 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" are obtained. Figure 2 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.

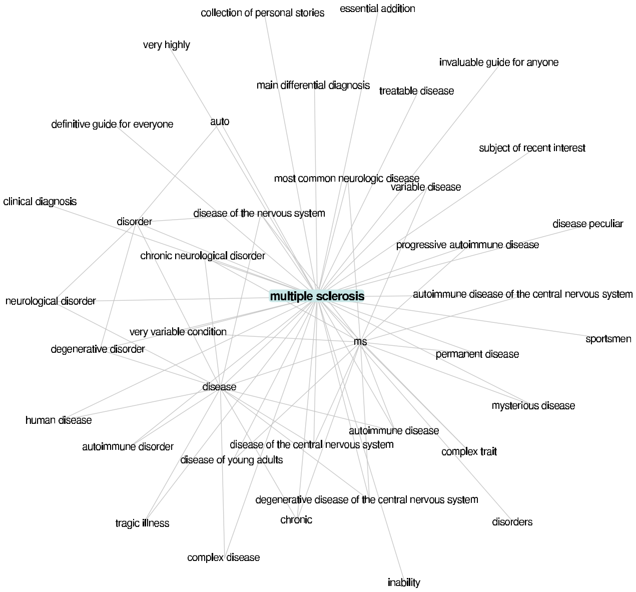
collection of personal stories
essential addition
very highly
invaluable guide for anyone
main differential diagnosis
treatable disease
definitive guide for everyone
auto
subject of recent interest
most common neurologic disease
variable disease
clinical diagnosis
disease of the nervous system
disorder
disease peculiar
progressive autoimmune disease
chronic neurological disorder
multiple sclerosis
autoimmune disease of the central nervous system
neurological disorder
very variable condition
ms
sportsmen
degenerative disorder
permanent disease
disease
mysterious disease
human disease
autoimmune disease
autoimmune disorder
disease of the central nervous system
complex trait
disease of young adults
degenerative disease of the central nervous system
tragic illness
chronic
disorders
complex disease
inability

**Figure 2. Hub concepts with degree $>=$ 8 for "Multiple Sclerosis".**

In general, these hub concepts relate to a variety of medical terms. A closer look on the concept *disease*, however, reveals that its usage is not exclusively limited to medicine. ConceptWorld collected a Web 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", or "antisemitism". Moreover, the range of concepts regarded as related to "disease" are captured with the following samples: "truth", "boredom", "terrorism", "poverty", "loss of spiritual equilibrium", "homosexuality", "war", "bigotry" or "aging". We refrain from showing the subgraph of the concept *disease* since the number of 227 neighbors would render the illustration unreadable.

The graph of the hub concept "disorder" identifies 83 concepts related to abnormal physical or mental conditions. 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" 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.

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. 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. An example from the information technology domain is shown by the results of a ConceptWorld process for “Microsoft 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” or “complicated system”, 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.

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”. 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, viz. AIDS is the autoimmune disease not HIV. Strictly speaking, this reflects people’s mistaken 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 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.

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.

## 4 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] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 279(5):34–43, 2001.

[2] R. Cilibrasi and P. Vitanyi. Automatic meaning discovery using google. Technical report, CWI, University of Amsterdam, 2004.

[3] L. Ding, T. Finin, A. Joshi, R. Pan, R. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: A search and metadata engine for the semantic web. In *Proc Int’l Conf on Information and Knowledge Management*, pages 652–659, Washington, D.C., USA, 2004. ACM Press.

[4] D. Fensel. The semantic web and its languages. *IEEE Intelligent Systems*, 15(6):67–73, 2000.

[5] M. Henzinger. Hyperlink analysis for the web. *IEEE Internet Computing*, 5(1):45–50, 2001.

[6] G. Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[7] R. McCool. Rethinking the semantic web, part 1. *IEEE Internet Computing*, 9(6):86–88, 2005.

[8] M. Schoop, A. de Moor, and J. Dietz. The pragmatic web: A manifesto. *Comm. of the ACM*, 49(5):75–76, 2006.