Ontology-Based Semantic Online Classification of Documents: Supporting Users in Searching the Web

Ernesto William De Luca and Andreas Nürnberger
Otto-von-Guericke University of Magdeburg
Universitätsplatz 2, 39106 Magdeburg, Germany
Phone: +49-391-67-18290, Fax: +49-391-67-12018
email: {deluca,nuernb}@iws.cs.uni-magdeburg.de

ABSTRACT. In this paper we describe first results of our research on the categorization of texts in order to disambiguate user queries. The discussed methods are based on a combination of indexing- and ontology-based information retrieval techniques in an interactive retrieval system. We present an approach to classify search results by mapping them to semantic classes that are defined by the senses of a query term. The criteria defining each class or 'sense folder' are derived from the concepts of an assigned ontology, here Multiwordnet. By annotating each element of a result set with the 'sense folder' it is classified in, the user gets additional information about each item – the specific search term is disambiguated with respect to the underlying document – and can thus decide more easily if the document is relevant or not for his query.

KEYWORDS: disambiguation, text classification, onotologies

INTRODUCTION

The amount of available information on the Internet has dramatically grown. At the moment, different search, indexing and cataloguing systems are easily accessible from the web, but the retrieval of relevant information (and also the management of it) is still limited. One current problem of information retrieval issues is that it is not really possible to automatically extract meaning from the relevant results of a query. One main reason for this is that the web was initially designed for direct human use [7] and thus the documents do not provide machine readable semantic annotations. However, if we analyze the situation of a user who interacts with a current information retrieval system, we come across the problem of understanding what the user means. If a human looks for information, he has an idea of what he is looking for, but he usually cannot formalize it – while using currently available search interfaces – in a form, so that a retrieval system can understand what is meant. What a user does is providing a set of keywords as a query in order to obtain information. Then, he is implicitly disambiguating his keywords while analyzing other content of the retrieved web pages in order to find the information he is looking for. In order to support a user in this process, we exploit the fact that every word conveys a certain meaning that can be made accessible and available using ontologies.

RELATED WORK

Several approaches have been presented in the past in order to classify or annotate documents. In the following we present very briefly some approaches that are using ontologies or hierarchies for this purpose. In [21] an approach was presented that uses Yahoo!-Categories as a concept hierarchy in order to classify documents using an n-gram classifier. The results presented in this paper indicate a reasonable performance of the system. However, a disadvantage of this approach is that it requires a predefined hierarchy and associated manually entered descriptions. Therefore, it can not be applied if the domain a user is interested in is not sufficiently covered or these descriptions are missing. Another problem is that the performance of the used classifier degrades if the hierarchy is too complex, i.e. too many categories have to be considered.

In [17] it was shown that searching using disambiguated terms can greatly improve the retrieval performance. Even for disambiguation errors of up to 30% the retrieval performance is better than that of standard term-weighting models. Even though, this study was based on WordNet annotated corpora, it shows that even imperfect disambiguation can strongly increase retrieval performance.

In [2] a query expansion system is described that uses a combination of automatic and user-assisted methods to build and improve cross-language queries. They try to use hyponyms, i.e. subordinate terms, in order to expand the query and improve the disambiguation process. However, refining the search by adding hyponyms to the original query usually removes too many documents from the result set, since hyponyms are very specific search terms. The problem the authors encounter in this specific setting is always concerning the quality and size of the ontology that strongly affect the results of the query.

SEMANTIC WEB AND ONTOLOGIES

With this vision of the Web as semantic network much more automated services based on machine-processable semantics of data are provided, and metadata are used to explicit represent the semantics of data accompanied with domain theories (i.e., ontologies). This provides a qualitatively new level of service concerning the "Knowledge Web". The concept of the "Semantic Web" given by Tim Berners-Lee [6] outlined the purpose of using machine-processable semantics of data and started a new evolution of the World Wide Web. This approach gives the possibility to sort and structure information in order to access and retrieve content precisely. A combination of semantical and statistical techniques makes it possible to retrieve more precisely documents that are linguistically related to the information searched [8, 12, 16]. This is provided by using ontologies in order to disambiguate the user query. However, one major problem we have to deal with is the fact that only a few web pages provide semantic annotations – and this will be a fact at least for the near future. Therefore, we decided not to rely on the annotation of documents, but to use different external resources to assign meaning to documents in relation to a given query in order to disambiguate their content as a user does when he is searching for information. Currently, people have to navigate among a lot of documents to discover the relevant one, because current retrieval systems do not provide such semantic information or relations.

An ontology is a formal, semantical specification of a conceptualization of a domain of interest. Ontologies are used to describe the semantics of information exchange. In natural language texts certain terms have different meanings that are not explicitly defined. Humans are able to disambiguate them by its context. However, current machine disambiguating approaches frequently fail due to missing commonsense knowledge or appropriate ontology models [15]. An important role for the disambiguation of the word context is the domain in which a word occurs. Linguistic ontologies can cover specific or general domains that are given as primitives. Primitives describe the generic terms that include other terms. An example is the primitive computer science that includes software, hardware, networks, etc. Therefore, ontologies such as Multiwordnet [23, 24, 25] can improve automatic disambiguation methods.

Multiwordnet is an expansion of (version 1.6) of WordNet[33, 14]. WordNet is an electronic lexical database that is considered to be the most important resource available to researchers in the field of text analysis, computational linguistics and many related areas. It was designed by the psycholinguistic and computational theories of human lexical memory. First it was developed only for the English language. Now it is also developed for several other languages. It contains descriptions of nouns, verbs, adjectives, and adverbs. These are organized into synonym sets (synsets), each representing one constitutional lexicalized concept. Every element of a synset is unambiguous and carrier of exactly one meaning. Furthermore, different relations link the elements of synonym sets to semantically related terms.

An ontology like Multiwordnet [5], can be used for a variety of content-based tasks, such as semantic query expansion or conceptual indexing in order to improve retrieval performance [10]. However, one major problem of query expansion is that the result set of documents is drastically reduced and therefore possibly relevant documents might get lost. Since query expansion has failed in several prior studies, mainly due to the fact that query expansion requires an almost 'perfect' set of keywords that is very hard to derive automatically, we decided to provide additional (structural) information to the user, instead of automatically restricting the result set. In the approach we present in the following, linguistic ontologies are used for content-based annotation, i.e. disambiguating semantic information (provided from the ontology) is added to a standard result list of a user query.

As a rule, linguistic ontologies are large scale lexical resources that cover most words of a language and have a hierarchical structure based on the relations between concepts. In contrast to offline text classification or clustering, e.g. as presented in [20, 29] where WordNet hypernyms are used to improve the performance of a learned text classifier, we cannot exploit any categorical information of the text documents. Since we are looking for a domain independent approach, the number of categories would be intractable large. Therefore, in our approach we combine a standard vector-space based retrieval system with an annotation method based on an onotology, currently Multiwordnet.

AN APPROACH FOR SEMANTIC CLASSIFICATION

The technique we used to provide information about ambiguities and resulting categories of search results is based on the following steps:

1. The user types his query (keywords).
2. These search terms are matched with ontologies whereby word vectors (prototypes) describing each semantic category are created.
3. Search results are indexed
4. Search results are classified to its sense folder
5. Retrieved documents are visualized

This process should simplify the search process by providing the user with explicit information about ambiguities and this enables him to easily retrieve the subset of documents he is looking for.

CREATION OF PROTOTYPES FOR SEMANTIC CLASSES

The system starts analysing the user query. For every search term we select from the used ontology the belonging synset, i.e. the set defining the different meanings of a term, and based on the relations of the ontology terms that describe the context in which each element of the synset is used. Based on this terms prototypical word vectors of the disambiguating classes are constructed.

In [4] an analysis concerning the use of WordNet relations for word sense disambiguation was presented. The authors evaluated different combinations of WordNet relations in order to disambiguate words using a conceptual density algorithm. They show that some relations such as meronymy (has-part relation) does not improve the performance as expected. Another problem they point out, is that in WordNet not all semantic relations are available for all words, which might result in significant classification problems, since one disambiguating class might be described more specific than another class.

Based on these studies we chose to classify documents using the hypernyms (the superordinate word), the hyponyms (the subordinate word) and the belonging glosses. We use the hypernymy relation because it describes the superordinate word of our search words (words that are more generic than the search word), and thus separating one meaning from another. We do not use only the superordinate hyperonym of our search word, but we extract all hyperonyms until we reach the primitive of the word. It means that we divide a category from another where they intersect. We also decided to use the hyponymy relation, because with this relation all subordinated words related to the query are included. Both of these relations describe the restricting context. We also chose to use the glosses (human readable description of the words) included in the Multiwordnet ontology, because they give a deeper description of the synsets elements by words that are frequently used in this specific semantic context.

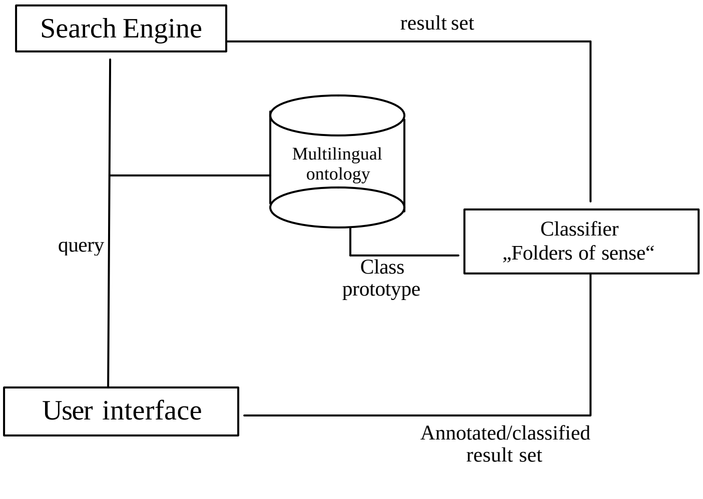


Figure 1: Overview of our system

We excluded the coordinate terms relation – the hyperonym of the word and the appropriate hyponyms of this hyperonym – because the obtained meanings of terms are too broad, i.e. many words would be included that did not resolve the problem of disambiguation since they are not closely related with the context.

We use all the words included in the relations mentioned above to define the vectors of the search context and thus for the different meanings of the user query. We call these vectors prototypes of the "sense folders".

In contrast to the approach presented in [4] that used only a small window of words around the considered term in order to disambiguate its meaning, we assume that the meaning of a search term used in a web-page can be defined based on the whole document, since web-pages are usually very short and usually cover only one semantic topic. This gives us a much better description of the context in order to disambiguate the search terms as described in the following.

INDEXING THE RESULT SET

After the vectors describing the disambiguating classes have been computed, the query is submitted to the search engine, which is then providing a set of search results. In order to be able to classify the search results, the returned documents have to be pre-processed. For the definition of the documents, we use a tf×idf-representation [28]. This means that each element in the vector corresponds to a term i in the document, while the size of the vector is defined by the number of words n occurring in the considered document collection (dictionary). The weights of the elements depend on term frequency $tf_{i,d}$ and inverse document frequency idf_i of the corresponding term. In our implementation, the weight w for the term i in the document vector of document d is calculated as follows (with N being the total number of documents and df_i being the document frequency of i):

$$w_{i,d} = \frac{\overline{w}_{i,d}}{\sqrt{\sum_{j=1}^{n} \overline{w}_{j,d}^2}} \text{ with } \overline{w}_{i,d} = tf_{i,d} \cdot \log\left(\frac{N+1}{df_i}\right)$$

Before vectors can be created, one has to decide which terms will be used to define the vector space. Usually, it is not reasonable to transfer every occurring term into the model. A high vector space dimensionality can have negative effects on computing time and quality of classification. In order to reduce the number of terms used for indexing, term selection and term extraction methods can be used. Term selection methods select a subset of – statistically or semantically relevant – terms for further use. We use stop word filtering and stemming methods (see e.g. [26]), which reduces each word to its stem., e.g. 'going' and 'goes' is mapped to 'go'.

DISAMBIGUATING THE SEARCH TERM

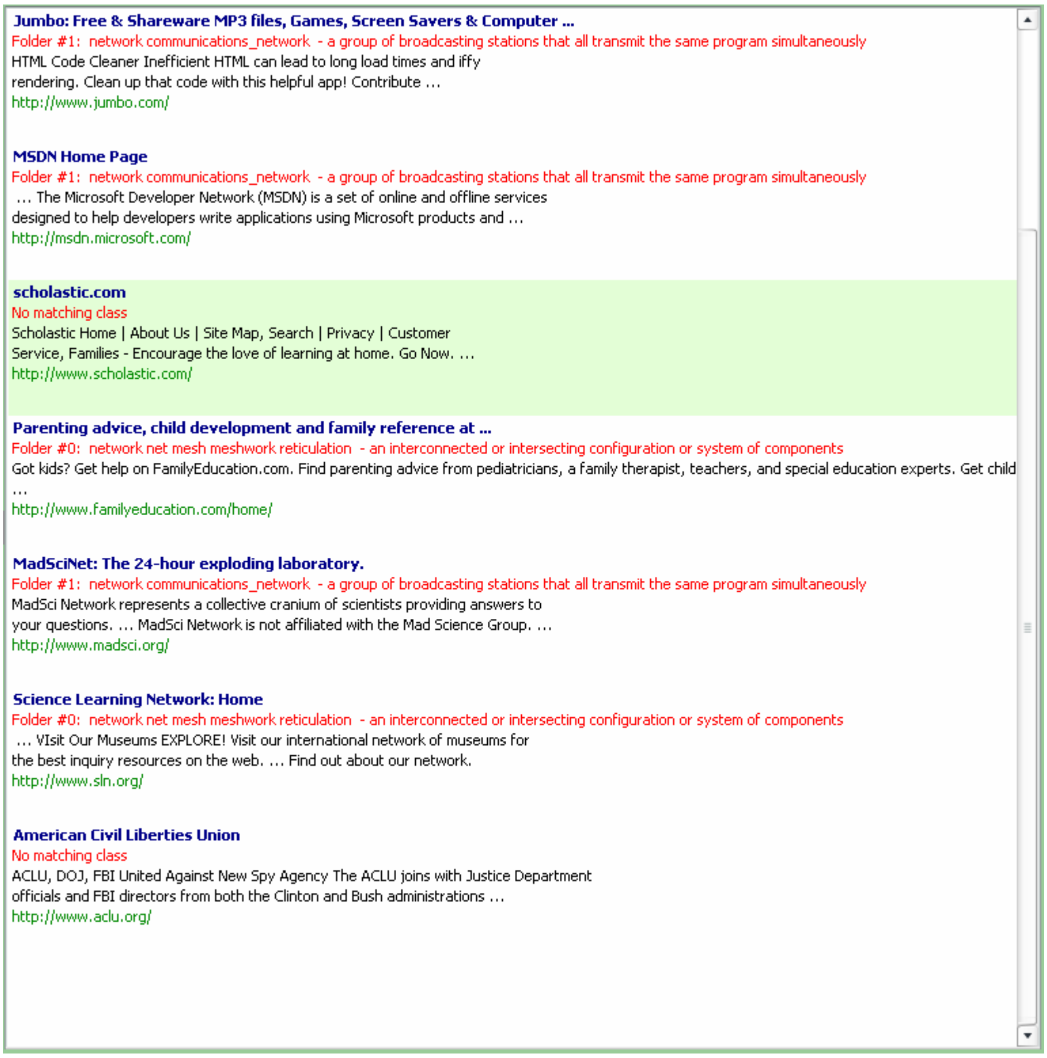
Once the vector space description for each element is computed, we classify documents by computing the similarity to each prototype vector describing the disambiguating classes (cosine similarity) and assigning the class with the highest similarity to the considered document.

If a user types, for example, the word "network", he has the possibility to obtain three different classification classes based on the noun collocations of this word (included in the Multiwordnet ontology) as shown in Table 1.

Table 1. Multiwordnet noun collocation of the term "network"

```
#0 network, web - an intricately connected system of things or people; "a network of
spies" or "a web of intrigue"
#1 network, communications_network - a group of broadcasting stations that all
   transmit the same program simultaneously
#2 network, net, mesh, meshwork, reticulation - an interconnected or intersecting
   configuration or system of components
```

The first collocation (#0) is related with the domain "factotum" and it has another context collocation as the others (e.g. the second one (#1) is related with the domain "telecommunication").



Jumbo: Free & Shareware MP3 files, Games, Screen Savers & Computer ...
Folder #1: network communications_network - a group of broadcasting stations that all transmit the same program simultaneously
HTML Code Cleaner Inefficient HTML can lead to long load times and iffy
rendering. Clean up that code with this helpful app! Contribute ...
http://www.jumbo.com/

MSDN Home Page
Folder #1: network communications_network - a group of broadcasting stations that all transmit the same program simultaneously
... The Microsoft Developer Network (MSDN) is a set of online and offline services
designed to help developers write applications using Microsoft products and ...
http://msdn.microsoft.com/

scholastic.com
No matching class
Scholastic Home | About Us | Site Map, Search | Privacy | Customer
Service, Families - Encourage the love of learning at home. Go Now. ...
http://www.scholastic.com/

Parenting advice, child development and family reference at ...
Folder #0: network net mesh meshwork reticulation - an interconnected or intersecting configuration or system of components
Got kids? Get help on FamilyEducation.com. Find parenting advice from pediatricians, a family therapist, teachers, and special education experts. Get child
...
http://www.familyeducation.com/home/

MadSciNet: The 24-hour exploding laboratory.
Folder #1: network communications_network - a group of broadcasting stations that all transmit the same program simultaneously
MadSci Network represents a collective cranium of scientists providing answers to
your questions. ... MadSci Network is not affiliated with the Mad Science Group. ...
http://www.madsci.org/

Science Learning Network: Home
Folder #0: network net mesh meshwork reticulation - an interconnected or intersecting configuration or system of components
... VIsit Our Museums EXPLORE! Visit our international network of museums for
the best inquiry resources on the web. ... Find out about our network.
http://www.sln.org/

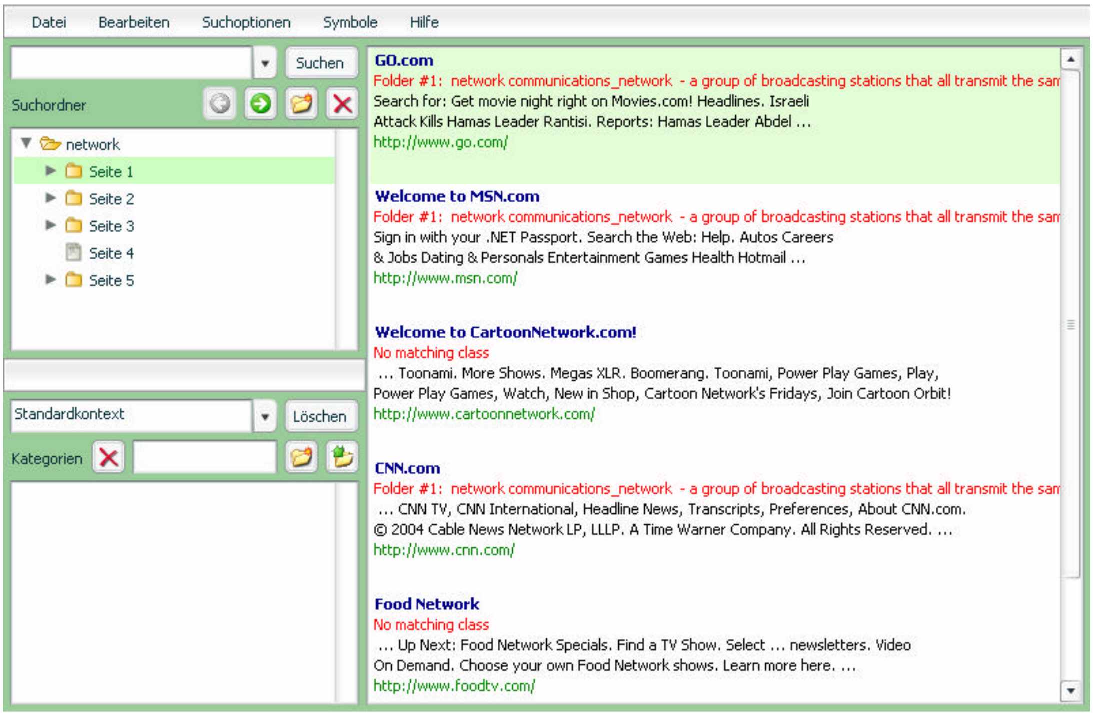
American Civil Liberties Union
No matching class
ACLU, DOJ, FBI United Against New Spy Agency The ACLU joins with Justice Department
officials and FBI directors from both the Clinton and Bush administrations ...
http://www.aclu.org/

Figure 2: Result of a query for 'network'

In Figure 2 the result of our classification of the result set for the typed search term "network" are shown. We can see that some results are classified in the right way according to the Multiwordnet category (b), as for example the first and the second document. The fourth and the sixth are classified to the category (c), whereby the fourth would not really match with the category at all. There are also other documents that are not assigned to any category, because are considered to not belong to any of the present categories.

IMPLEMENTATION

The current implementation realizes am information retrieval system that can be connected to any search engine that provides web services. Thus we are able to use the web coverage of the underlying search engine – currently Google – with the semantic online classification provided by our approach. Furthermore, the use of standardized ontologies as Wordnet (resp. Multiwordnet) gives us the possibility to develop a search engine that is language independent and upgradeable (we can add more or refined ontologies in order to improve the precision of our results).



GO.com
Folder #1: network communications_network - a group of broadcasting stations that all transmit the sam
Search for: Get movie night right on Movies.com! Headlines. Israeli
Attack Kills Hamas Leader Rantisi. Reports: Hamas Leader Abdel ...
http://www.go.com/

Welcome to MSN.com
Folder #1: network communications_network - a group of broadcasting stations that all transmit the sam
Sign in with your .NET Passport. Search the Web: Help. Autos Careers
& Jobs Dating & Personals Entertainment Games Health Hotmail ...
http://www.msn.com/

Welcome to CartoonNetwork.com!
No matching class
... Toonami. More Shows. Megas XLR. Boomerang. Toonami, Power Play Games, Play,
Power Play Games, Watch, New in Shop, Cartoon Network's Fridays, Join Cartoon Orbit!
http://www.cartoonnetwork.com/

CNN.com
Folder #1: network communications_network - a group of broadcasting stations that all transmit the sam
... CNN TV, CNN International, Headline News, Transcripts, Preferences, About CNN.com.
© 2004 Cable News Network LP, LLLP. A Time Warner Company. All Rights Reserved. ...
http://www.cnn.com/

Food Network
No matching class
... Up Next: Food Network Specials. Find a TV Show. Select ... newsletters. Video
On Demand. Choose your own Food Network shows. Learn more here. ...
http://www.foodtv.com/

Figure 3: Screenshot of the user interface

Figure 3 shows a screenshot of the user interface of our system. The interface basically consists of three parts, the query area, the contexts and categories area, and the result list area. The query area can be found on the upper left side. Here, the user can enter his queries or can scan through the results from the current or previous queries by selecting a specific result list. As all previous queries performed by the user and the corresponding results are stored by the interface, the user is able to switch arbitrarily between the results.

A context and categories area is located underneath the query area. This area enables a user to store links (bookmark concept). The user can browse through it, create and delete contexts and categories in a hierarchical tree structure, assign single web sites to certain categories and view the web sites contained in each category. There is always one context selected as the current one.

The result list area is on the right side. Here, a list of web sites is presented. This can either be a list of search results or a list of web sites contained in a selected category. Each web site is represented by its title, its hyperlink, and a snippet as it is provided by standard search engines. In addition, the disambiguating category as determined by our algorithm is presented for each search result item.

FIRST EVALUATION RESULTS

In the following we present the results of a first system evaluation. For that purpose we created a test set by manually annotating the first fifty Google hits (as of 02.02.2004) for the term “network” with their correct meaning – as defined by the Multiwordnet ontology – in the context of the document. Afterwards the same task was given to the system. The three word-sense classes (#0, #1, #2) are those described in Table 1. The “no class” entry denotes a meaning that is – in our opinion – not appropriately included in the ontology. The categorization algorithms return “no class” if none of the keywords of the prototypes can be found in the web site (in this case the cosine-similarity of the document to all prototypes is zero). The automatically generated classification was then compared to the manual classification. The results are shown in the confusion matrix in Table 2. In case of a perfect classification, only the main diagonal should be filled. As can be seen from the following matrix our classification shows promising results. In the specific case of the query “network”, we obtained the desired result of 68% of the cases. However, in order to obtain significant results, evaluations with different search terms and larger result sets have to be performed.

Table 2. Confusion matrix of system evaluation (searching for "network")

Correct Class / Predicted Class	**0**	**1**	**2**	**No class**
0	12	2	2	1
1	1	14	0	5
2	1	0	2	1
No class	1	2	0	6

FUTURE WORK

The performance of the currently integrated disambiguation method still needs to be improved. Therefore, different ontologies and more refined semantic classification techniques have to be studied. Furthermore, a more detailed analysis of the performance of different concept combinations that are extracted from the ontology based on their relation to the search term is necessary.

A long term goal of our work is to develop a multilingual information retrieval system. Therefore, we are confronted with different issues that have to be considered. We can assume that in the majority of the cases a word representation describes the same object or concept, also when it is described in different languages. Naturally there are also cases where such a word representation has no comparable translation or the word does not exist in other languages. Therefore, a similar disambiguating process should be performed for the translated terms. This could be done by integrating different ontologies in our system. To merge two ontologies a common point between concepts has to be found (for example, in Multiwordnet using synset ids that are language independent). Our goal is to create such a multilingual system that offers the possibility to perform the retrieval by mapping, for example, synsets (it means more collocations from the ontology) to an interlingual index that includes all ontologies that can be monolingual, bilingual or multilingual. Thus, if a user is, for example, searching for information in English and he does not get any relevant hits or documents, he should be able to get automatically results in a different language. The results should be retrieved in parallel for as many languages as there are available in the ontology. This form of query adaptation of our multilingual search system will offer the possibility to query the web and get results classified in meaning and language.

Another important concept for a retrieval system is the adaptation to user needs. In order to provide an individual support to the user, adaptive techniques must be integrated. Our system is designed to interactively support the user in his query and currently different techniques for classification and clustering of result sets are integrated that consider user specific interests [34].

CONCLUSIONS

In this paper we have presented a search system that uses ontologies to classify search results online in order to disambiguate result sets with respect to given search terms. Thus, the user can select directly a subset of the search results ("folder of sense") which reflects his search context without the need to scan the list of all retrieved documents. So far, the results of our approach are very promising. However, a more detailed analysis of the performance of the disambiguation approach is necessary.

The resulting visualization of the semantic categories is query and not collection oriented. Our goal is to develop an upgradable information retrieval system using different ontologies and merging different search procedures having at the end a multilingual retrieval system.

ACKNOWLEDGEMENTS

The work presented in this article was supported by the German Research Community (Deutsche Forschungsgemeinschaft, DFG), project number NU 131/1-1 and the European Network of Excellence on Intelligent Technologies for Smart Adaptive Systems (EUNITE).

REFERENCES

[1] Adaptive Hypertext and Hypermedia Home Page, http://wwwis.win.tue.nl/ah/.

[2] Ahmed Abdelali, James Cowie, David Farwell, Bill Ogden, and Stephen Helmreich, "Cross-Language Information Retrieval using Ontology", Proceedings of TALN 2003, Batz-sur-Mer, France.

[3] Stuart Aitken, and Sandy Reid, "Evaluation of an Ontology-Based Information Retrieval Tool", Proceedings in ECAI 2000, Berlin, Germany.

[4] E. Agirre and G. Rigau. 1996. Word sense disambiguation using conceptual density. In Proceedings of COLING'96, pages 16--22, Copenhagen, Danmark.

[5] Luisa Bentivogli, Emanuele Pianta, Christian Girardi, 2002, "MultiWordNet: developing an aligned multilingual database", Proceedings of the First International Conference on Global WordNet, Mysore, India.

[6] Tim Berners-Lee, 1998, "Semantic Web Road map", http://www.w3.org/DesignIssues/Semantic.html

[7] P. Brusilovsky, 1996, "Methods and techniques of adaptive hypermedia", in User Modeling and User Adapted Interaction, v.6, n.2-3.

[8] P. Brusilovsky, A. Kobsa, J. Vassileva (eds.), 1998, "Adaptive Hypertext and Hypermedia", Kluwer Academic Publishers.

[9] P. Brusilovsky, O. Stock, C. Strapparava (eds.), 2000, "Adaptive Hypermedia and Adaptive Web-Based Systems", Proceedings of the International Conference AH 2000, Trento, Italy.

[10] Fabio Ciravegna, Bernardo Magnini, Emanuele Pianta, Carlo Strapparava, 1994, "Multilingual Lexical Knowledge Bases: Applied WordNet Prospects", International Workshop on "The Future of the Dictionary", Grenoble.

[11] Fabio Ciravegna, Bernardo Magnini, Emanuele Pianta, Carlo Strapparava, 1994, "A Project for the Construction of an Italian Lexical Knowledge Base in the Framework of WordNet". IRST Technical Report # 9406-15.

[12] Ying Ding, Cornelis J. van Rijsbergen, Iadh Ounis, Joemon Jose, 2003, "Report on ACM SIGIR Workshop on 'Semantic Web' SWIR 2003". Toronto, Canada.

[13] M. Erdmann, A. Maedche, H.-P. Schnurr, S. Staab, "From Manual to Semi-automatic Semantic Annotation: About Ontology-based Text Annotation Tools"

[14] Christiane Fellbaum, WordNet, an electronical lexical database. 1998, Cambridge, MA: MIT Press.

[15] Dieter Fensel, Frank van Harmelen, Michel Klein, Hans Akkermans et al. , "On-To-Knowledge: Ontology-based Tools for Knowledge Management", Free University Amsterdam, The Netherlands, 2000.

[16] J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarrán, Indexing with WordNet synsets can improve text retrieval, in Proc. of the COLING/ACL '98 Workshop on Usage of WordNet for NLP

[17] Julio Gonzalo Felisa Verdejo Irina Chugur Juan Cigarrán , "Indexing with WordNet synsets can improve text retrieval". Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP, Montréal/Canada.

[18] B. Hayes, The web of words. In American Scientist Volume 87 Number 2, pages 108-112, 1999.

[19] Eero Hyvönen, Samppa Saarela, and Kim Viljanen, "Ontogator: Combining View- and Ontology-Based Search with Semantic Browsing", Proceedings of XML Finland 2003, Open Standards, XML, and the public Sector, Kupio, 2003.

[20] Andreas Hotho, Steffen Staab, Gerd Stumme, 2003, "Wordnet improves Text Document Clustering". ACM SIGIR Workshop on 'Semantic Web', Toronto, Canada.

[21] Yannis Labrou and Tim Finin, “Yahoo! as an ontology: using Yahoo! categories to describe documents”. Proceedings of the eighth international conference on Information and Knowledge Management, Kansas City, Mis souri, 1999.

[22] Bernardo Magnini, Manuela Speranza, 2002. ”Merging Global and Specialized Linguistic Ontologies”, Proceedings of the Workshop Ontolex-2002 Ontologies and Lexical Knowledge Bases, LREC-2002, pp. 43-48.

[23] Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, Alfio Gliozzo, 2002, ”Comparing Ontology-Based and Corpus-Based Domain Annotation in WordNet”, Proceedings of First International WordNet Conference, Mysore, India, pp. 146-154.

[24] Bernardo Magnini, Carlo Strapparava, Fabio Ciravegna and Emanuele Pianta, 1994, “A Project for the Construction of an Italian Lexical Knowledge Base in the Framework of WordNet”. IRST Technical Report # 9406-15.

[25] Emanuele Pianta, Luisa Bentivoglio, Christian Girardi, 2002, “Multiwordnet: Developing an Aligned Multilingual Database”. In Proceedings of the First International Conference on Global WordNet, Mysore, India, pp. 293-302.

[26] Porter, M., 1980, “An algorithm for suffix stripping”, Program, pp. 130-137.

[27] Vilho Raatikka, and Eero Hyvönen, “Ontology-based Semantic Metadata Validation”, University of Helsinki, Department of Computer Science

[28] Salton, G.; Buckley, C., 1988, “Term Weighting Approaches in Automatic Text Retrieval”, Information Processing & Management, 24(5), pp. 513-523.

[29] Sam Scott, Stan Matwin, 1999, “Feature Engineering for text classification”. In Ivan Bratko and Saso Dzeroski, editors, Proceedings of ICML-99, 16th International Conference on Machine Learning, San Francisco, pages 379-388.

[30] Sam Scott, Stan Matwin, 1998, Text Classification Using WordNet Hypernyms. . In Proc. of the Conf. on Use of WordNet in Natural Language Processing Systems, ACL, 1998, 38-44

[31] Anna Stefani and Carlo Strapparava, 1999, “Exploiting NLP techniques to build user model for Web sites: the use of WordNet in SiteIF Project”. In P. Brusilovsky, P. De Bra & A. Kobsa (Chairs), Proceedings of the Second Workshop on Adaptive Systems and User Modeling on the WWW.

[32] Carlo Strapparava, Bernardo Magnini and Anna Stefani, 2000, “Word Sense Based User Models for Web Sites”. 9th International World Wide Web, The Web: The Next Generation. Amsterdam.

[33] Wordnet homepage, http://www.cogsci.princeton.edu/~wn/ („Five Papers on Wordnet“)

[34] A. Nürnberger und M. Detyniecki, 2002, “User Adaptive Methods for Interactive Analysis of Document Databases”, In *Proc. of the European Symposium on Intelligent Technologies (EUNITE 2002)*, Albufeira, Portugal, Verlag Mainz, Aachen.