Improving Ontology-Based Sense Folder Classification of Document Collections with Clustering Methods

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
{deluca,nuernb}@iws.cs.uni-magdeburg.de

Abstract. In this paper we describe first results of our research on the disambiguation of user queries using ontologies for categorization. We present an approach to cluster search results by using classes or 'Sense Folders' (prototype categories) derived from the concepts of an assigned ontology, here MultiWordNet. Using the semantic relations provided from such a resource, we can assign categories to prior not annotated documents and thus help the user in finding the relevant documents related to his search keywords. Furthermore, we show that a clustering process can enhance the semantic classification of documents.

1 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 systems is that it is not really possible to automatically extract meaning from the relevant results of a query in order to support the user in the search process. One main reason for this is that the web was initially designed for direct human use and thus the documents do not provide machine readable semantic annotations.

The concept of the "Semantic Web" proposed by Tim Berners-Lee [3] outlined the idea of using machine-processable semantics assigned to each available information resource and started a new evolution of the World Wide Web. This approach should provide the possibility to sort and structure information in order to access and retrieve content more precise. This vision of the Web represents the idea to "migrate" to a "Knowledge Web", where all concepts are organized as a semantic network accompanied with domain theories (i.e., ontologies). Thus, this approach provides the possibility to combine semantical and statistical techniques in order to retrieve documents that are linguistically related to the information a user is searching for more precisely [4, 7, 12].

However, given that there are currently only a few web pages that provide semantic annotations, we decided to use only the already available external resources (in our case ontologies) to assign meaning to documents in relation to a given query. The

idea is to disambiguate their content similar to a user who is searching for information. Currently, people have to navigate among a lot of documents to discover this disambiguating information on their own in order to select the relevant documents, because current retrieval systems do not provide such semantic information or relations.

2 Use of Ontologies in Information Retrieval Systems

An ontology is a formal specification of a conceptualization of a domain of interest. It specifies a set of constraints, which declare what should necessarily hold in any possible world. Ontologies are used to identify what "is" or "can be" in the world. It is the intention to build a complete world model for describing the semantics of information exchange. Especially in the area of artificial intelligence ontologies are being used in order to facilitate knowledge sharing and reuse.

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 [9]. An important role for the disambiguation of the word context is the domain in which a word occurs.

Currently, there is still no ontology available that can directly be used to disambiguate meanings. A rigorous but clear logical representation of the concepts of the world is needed to increase their transparency and interoperability. If this representation is good enough, a better disambiguation process could be feasible. Research in the field of philosophy and linguistics is currently working in this direction [10].

In general, 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. These 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 Wordnet [8, 23] or its variant MultiWordNet [14, 17] can improve automatic disambiguation methods.

2.1 Wordnet

Figure 1 represents an example of the ontology hierarchy defined by WordNet. WordNet is one of the most important resource available to researchers in the field of text analysis, computational linguistics and many related areas. It is an electronic lexical database designed by use of psycholinguistic and computational theories of human lexical memory. It provides a list of word senses for each word, organized into synonym sets (Synsets), each representing one constitutional lexicalized concept. Every element of a Synset is uniquely identified by its Synset identifier (SynsetID). It is unambiguous and carrier of exactly one meaning. Furthermore, different relations link these elements of synonym sets to semantically related terms (e.g. hyperonyms, hyponyms, etc.). All related terms are also represented as Synset entries.

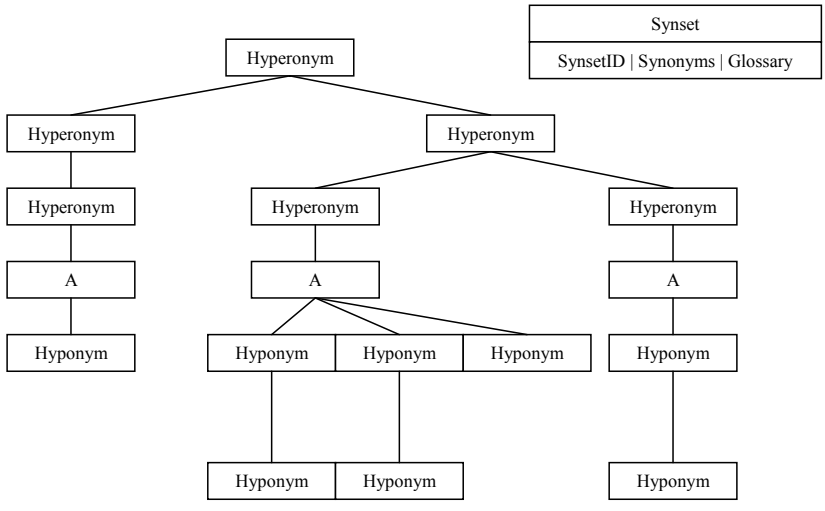

Fig. 1 Example of an ontology hierarchy for a given term A

WordNet also contains descriptions of nouns, verbs, adjectives, and adverbs. First it was developed only for the English language. The current version of MultiWordNet [2] is an expansion of (version 1.6) of WordNet and it contains English, Italian, and Spanish terms. Also other versions of WordNet are being developed for several other languages (e.g. EuroWordnet [20] for almost all European languages).

3 Categorization of documents using ontologies

An ontology like WordNet, can be used for a variety of content-based tasks, such as semantic query expansion or conceptual indexing in order to improve retrieval performance [5, 14]. In prior work, we have studied the use of ontologies in order to disambiguate terms used in documents with respect to a given user query [6]. In the approach we present in the following, we combine this approach with clustering methods in order to improve the quality of the disambiguation process.

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 (see Sect. 3.1).
3. Search results are indexed (see Sect. 3.2).
4. Search results are classified to its Sense Folder (see Sect. 3.3).
5. The result set is clustered using Sense Folders as initial centers for the clustering process (see Sect. 3.4).

These processing steps have been integrated in our document retrieval system. Figure 2 shows its structure and the data flow within the system. The process starts when the user types his query. The system analyses every search term and extracts the belonging Synsets, i.e. the sets defining the different meanings of a term and the linguistic relations from the used ontology. Based on this terms prototypical word vectors of the disambiguating classes are constructed. Every document is assigned to its nearest prototype and afterwards this classification is revised by the clustering process. The processing steps are described in more detail in the following. For a more detailed description of the system architecture see [6].

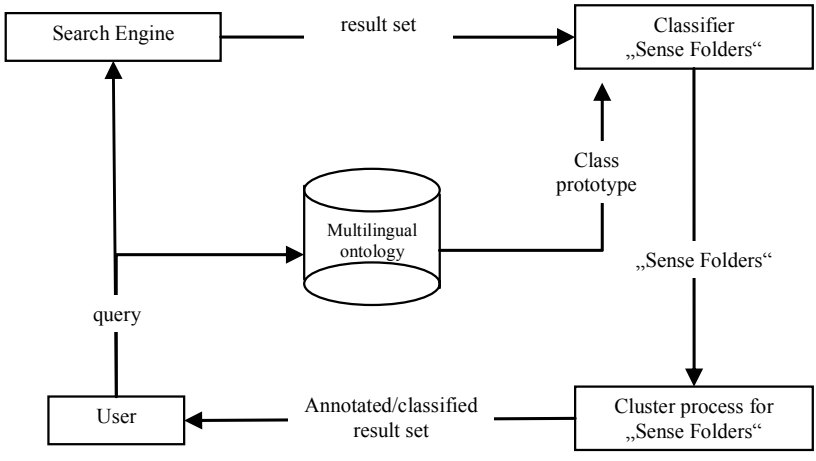

Fig. 2 Overview of the retrieval system

3.1 The use of linguistic relations for creating prototypes of semantic classes

In order to assign documents to the semantic class they belong to, we have to consider the different linguistic relations that describe the context of the searched word in the ontology. We chose to classify documents using 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 WordNet ontology, because they give a deeper description of the Synset elements by words that are frequently used in this specific semantic 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” [6]. In contrast to the approach presented in [1] 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.

3.2 Indexing the results set

After the vectors describing the disambiguating classes have been computed, the query is submitted to the search engine, which is then retrieving a set of documents that match the query.

Since we are dealing with natural language text, the documents have to be pre-processed in order to be able to store their information in a data structure that is more appropriate for classification and clustering. Therefore, we encoded each document by the commonly used vector space model [19]. The vector space model represents text documents as vectors in an *m*-dimensional space, i.e., each document is described by a numerical feature vector. Each element of the vector represents a word of the document collection, i.e., the size of the vector is defined by the number of words of the complete document collection.

A high vector space dimensionality can negatively influence the computation time required and the quality of the classification [13]. In order to reduce the number of terms used for indexing, term selection and term extraction methods can be used. Thus, before describing vectors for each document can be created, one has to decide which terms should be used to define the vector space. Usually, it is not reasonable to transfer every occurring term into the model. In the current implementation, term selection is done by stop word filtering and stemming methods (see e.g. [18]), which reduces each word to its stem., e.g. ‘going’ and ‘goes’ is mapped to ‘go’. Finally, we compute a tf×idf-representation for each document [20].

3.3 Disambiguating the meaning of the search terms

In our current implementation, we are using the cosine similarity measure for document classification. Once the vector space description for each document is computed, we classify the documents by computing the similarity to each prototype vector describing the disambiguating classes and assigning the class with the highest similarity to the considered document.

If a user types, for example, the word “range”, he has the possibility to obtain nine different classification classes based on the noun collocations of this word (included in the WordNet ontology) as shown in Table 1. Here, the first Synset (#0) is related with the domain “Factotum” and it describes another context or meaning than the other Synsets (e.g. the second one (#4) is related with the domain “geology”).

Table 1. WordNet noun collocation of the term "range"

#0 scope, range, reach, orbit, compass, ambit	(Factotum) an area in which something acts or operates or has power or control: "the range of a supersonic jet"; "the ambit of municipal legislation"; "within the compass of this article"; within the scope of an investigation"; "outside the reach of the law"; "in the political orbit of a world power"
#1 range, reach	(Factotum) the limits within which something can be effective; "he was beyond the range of their fire"
#2 range	(Mathematics) the limits of the values a function can take; "the range of this function is the interval from 0 to 1"
#3 range	(Agriculture) a large tract of grassy open land on which livestock can graze; "they used to drive the cattle across the open range every spring"; "he dreamed of a home on the range"
#4 range, mountain_range, range_of_mountains, chain, mountain_chain, chain_of_mountains	(Geology) a series of hills or mountains; "the valley was between two ranges of hills"; "the plains lay just beyond the mountain range"
#5 range	(Factotum) a place for shooting (firing or driving) projectiles of various kinds; "the army maintains a missile range in the desert"; "any good golf club will have a range"
#6 range	(Factotum) a variety of different things or activities; "he answered a range of questions"; "he was impressed by the range and diversity of the collection"
#7 compass, range, reach, grasp	(Factotum) the limit of capability; "within the compass of education"
#8 stove, kitchen_stove, range, kitchen_range, cooking_stove	(Furniture) a kitchen appliance used for cooking food; "dinner was already on the stove"

3.4 Clustering the Sense Folders

After the document vectors are assigned to their respective WordNet class, we use the *k*-means clustering algorithm to tune our classification results. We use the number of classes obtained from the ontology for the number of clusters *k* and start the clustering process using the WordNet Sense Folders as initial cluster centers.

The *k*-means algorithm computes the cluster centers by simply repeating two steps. First, each object is assigned to the closest cluster center. Then the cluster centers are recalculated as the average of the document vectors assigned to the cluster. This process is repeated until there are no more changes to the cluster centers. A pseudocode description of the implemented algorithm is shown in Figure 3.

```
K-Means Algorithm(k, Sense Folders)
   Initialize the k clusters
         Cluster centres are set to the WordNet
         Sense Folders
   Repeat
         For Each Document Vector
            Assign to the Cluster with the closest
            prototype
         For Each cluster
            Calculate new Cluster centre as average of
            document vectors assigned to the cluster
   Until no more reallocations of documents occur
```

Fig. 3. Implementation of the k-means algorithm

For our disambiguation problem, we use the clustering algorithm in order to fine tune the initial prototype vectors of each Sense Folder using the distribution of documents around the initial prototype vectors, i.e., we expect that in a web search usually a subset of documents for each possible meaning of a search term is retrieved. Thus, each subset forms a cluster in document space describing one semantic meaning of this term.

4 Implementation and empirical evaluation results

We have implemented the methods described above in the retrieval system described in [6]. The retrieval system is realized as a Meta-Search engine and thus is able to combine the web coverage of the underlying search engine – currently Google – with the semantic online classification provided by the approach described above. Furthermore, the use of standardized ontologies as WordNet 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). In the following we present results of empirical evaluation and discuss some problems of WordNet based disambiguation.

4.1 Disambiguation of the result set of a web search

In Table 2 a subset of the search results for the term “range” is depicted. The subset represents the documents that are assigned by the clustering process (see Sect. 3.4) to the two clusters #3 and #5 (first column of Table 2). The cluster #3 considers the word “range” in the context of “Agriculture” while the cluster #5 considers it in the context of “Factotum” in the sense of shooting. The links to the documents retrieved from Google are listed in the third column. The classification using the pure Sense Folder based disambiguation process (see Sect. 3.1 and 3.3) is shown in the second column of Table 2 (see Table 1 for a description of the belonging WordNet Synsets).

We include also a “no class” (“Sense Folder -1”) entry that denotes that the prototype vectors created from the ontology do not match any word in the document, i.e.

the scalar product between all Sense Folder prototypes and the document is zero. The reason for this could be that the meaning is not appropriately included in the ontology or that the Synset entry of the ontology provides too few describing terms. Especially for these cases, the clustering process can be beneficial, since it groups documents based on all terms that the documents share. Thus a document that contains no keyword from the Sense Folder prototypes might as well share several terms with documents describing a similar context and can therefore be assigned to the same class.

In the following, we discuss briefly based on some examples the improvements obtained by the clustering process. A more quantitative evaluation is given in Sect. 4.3.

First, we can observe that the Sense Folders that were classified as "no class" during the disambiguation process using the ontology (in Table 2 they are marked as "Sense Folder -1") are now – with the clustering process – in two cases classified to the correct class.

Furthermore, in several cases the clustering process successfully re-grouped documents that belong to the same context (having similar words) together. This is especially remarkable for cluster #5, where the pure Sense Folder based approached assigned most documents initially to the wrong Synset meanings. Here the clustering approach was able to reassign several documents to the correct class. This is depicted in Table 2 by the underlined documents are that successfully reclassified and assigned for example from the Sense Folders #0 and #8 to the Cluster #5 or #3.

The documents highlighted in boldface belong to none of the meaning classes defined by the ontology. Here, one of the most difficult tasks in disambiguation is emphasized. If a user is querying using the term "range" the search engine retrieves all documents that contain this term, without knowing what their real content is. The task of the ontology is to provide all possible meanings (Sense Folders) a document could be classified with. But, for example, the documents marked in Table 2 in boldface do not have anything in common with the meaning of "range". They belong to the class "car" that has an instance called "Range Rover" which is not described by the used WordNet ontology.

This sort of problem points out one main problem of the use of ontologies in information retrieval systems: the use of a general ontology like WordNet is not always recommended. If the search can be restricted to a specific topic, we should use a more specific ontology (e.g. a domain ontology) in order to restrict the context we work with. Thus it is possible to provide better disambiguation results and the meaning or context can be described much clearer to the user.

However, some problems mentioned above could be avoided if WordNet is restructured for disambiguation purposes as briefly discussed in the following.

Table 2. Subset of a classified result set for keyword "range" (Underlined = documents successfully reclassified with clustering, **Bold** = documents not belonging to any ontology meaning, Normal = document class not changed, *Italic* = wrong cluster assignment)

Cluster	*Sense Folder*	*Document link*
3	8	http://www.irrrb.org/
3	0	http://www.rw.ttu.edu/dept/
3	-1	*http://robotics.jpl.nasa.gov/tasks/scirover/homepage.html*
3	3	**http://www.landrover.com/**
3	0	**http://www.carsource.co.uk/**
3	0	**http://www.4x4web.co.uk/**
3	-1	http://robotics.jpl.nasa.gov/tasks/scirover/
3	8	http://www.uidaho.edu/cfwr/
3	4	**http://www.carelect.demon.co.uk/rrind.html**
3	8	**http://www.rimmerbros.co.uk/**
3	8	http://extension.usu.edu/rangeplants/
3	8	http://www.aves.net/ohiomaps/
3	3	http://www.ksu.edu/weather/rfd.html
3	6	*http://www.cpc.ncep.noaa.gov/products/forecasts/*
3	3	*http://www.cpc.ncep.noaa.gov/products/predictions/*
3	3	http://wstc.wsmr.army.mil/capabilities/range_inst.html
3	-1	http://www.agronomy.ucdavis.edu/agronomy/default.htm
3	0	http://aec.army.mil/usaec/range/sustainment00.html
3	8	http://www.hill.af.mil/uttr/range.htm
5	8	http://www.vdkpublishing.com/
5	1	http://www.de.afrl.af.mil/
5	0	http://www.redsguns.com/
5	4	http://www.foprangeinc.com/
5	4	http://www.targetmaster.com/
5	0	http://www.lrml.org/
5	0	http://www.caswells.com/
5	8	http://www.therangeinc.com/
5	0	http://www.nrahq.org/shootingrange/nrahqrange/
5	8	http://www.handhgunrange.com/
5	0	http://www.nasr.com/
5	0	http://www.sevierindoorrange.com/
5	8	http://www.ca.blm.gov/arcata/birdlist.html
5	8	http://www.packing.org/range/
5	5	http://www.snailtraps.com/
5	0	http://www.laxrange.com/

4.2 Restructuring WordNet Synsets

Similar to [10, 11, 16] we examine in the following briefly the main semantic limitations of Wordnet (resp. MultiWordNet) and describe some problems that have to be solved for a better expressiveness of WordNet.

Some lexical links of Wordnet should be interpreted using formal semantics in order to express "things in the world". The authors of [16] revise the Top Level of Wordnet (upper or general level) where the criteria of identity and unity are very general, in order to recognize the constraint violations occurring in it. (The concepts of identity and unity are described in [16].) However, we analyze the expressiveness of every Synset in order to better categorize the context for clustering purposes. It means that we merge categories that are in the same domain and that are not much different from another. This decision is based on our need of few unique classes that are carrier of an expressive meaning for a user as well as for an improved clustering performance.

An example is given in Table 2 for the word "rule". If we retrieve 'rule' from Wordnet, we get 12 different meanings of this term. Several meanings are assigned to the domain "Factotum" that could be described as the class "other domain, generic". The reason for this assignment is simply the problem that the Wordnet authors have to assign a domain to each Synset. If a term can not be categorized (by the author) to a more specific domain, the generic domain "Factotum" is used. Therefore, if we want to categorize documents with WordNet senses, we have to choose which senses are relevant and which are not, in order to obtain appropriate disambiguation results. However, if we maintain all senses that are labeled with "Factotum", we have in many cases to distinguish between only slightly different contexts defined by different Synsets. One possibility to derive terms that have a very similar meaning is to analyze their hyponyms or hyperonyms. If there are two senses described in WordNet belonging to the same domain, they often have the same hyponyms or hyperonym. This frequently causes disambiguation problems that can not be solved if we keep all classes. For this reason, we decided to exclude some irrelevant (for the context disambiguation process) "Factotum" Synsets.

Another critical point is given by the confusion between concepts and instances resulting in an "expressivity lack" [11]. For example, if we look for the hyponyms of "mountain" in WordNet, we will find the "Olympus mount" as a subsumed concept of the word treated as "volcano" and not as instance of it. Thus, we do not have a clear differentiation between what we use to describe (concepts) and their instantiation (instances). We also have the problem that we can not use only concepts or only instances because there is no intended separation between them in Wordnet.

If we stay within the "rule" example, we can see that there are two categories that are labeled with the same words "principle rule" in the domain "Factotum". There are also two other categories that are not really relevant for disambiguating the senses, as for example "rule linguistic rule" and "rule regulation" that should be considered as instances of "rule". Referring to our discussion above, we excluded the Synsets #2, #9, #11 because they define instances and not concepts. Furthermore, we merged the classes #4 and #5 because they describe almost the same concept as "principle, rule", and the classes #0 and #3 because they belong both to mathematics. The Synsets #6,

#7, #10 are labeled with the "Factotum" entry and are not very expressive, but too general to be joined. Therefore, we removed them.

The authors of [15] treat also the important difference between endurance and perdurance of the entities that should be included in Wordnet. Enduring and perduring entities are related to their behavior in time. Endurants are always wholly present at any time they are present. Perdurants are only partially present, in the sense that some of their proper parts (e.g., their previous phases) may be not present. However, these aspects of instances are not discussed in this paper since they seem to be of less importance for the considered disambiguation problem.

In the following, we present an evaluation of a restructured Synset of the term "rule" for the categorization of the documents in a labelled benchmark dataset.

Table 2. Comparison of WordNet Synsets and restructured Synsets for clustering

Wordnet Synset	Restructured Synset
#0 rule ruler (Metrology)	#1
#1 convention normal pattern rule formula (Sociology)	#0
#2 rule regulation (Factotum – behavior)	
#3 rule formula (Mathematics)	#1
#4 principle rule (Factotum – rule, law)	#2
#5 principle rule (Factotum – generalization)	#2
#6 rule (Factotum – religion)	
#7 rule prescript (Factotum – guide)	
#8 rule (Factotum – game, sport)	#3
#9 rule linguistic_rule (Linguistics)	
#10 dominion rule (Factotum – legal authority)	
#11 rule (History Time_Period)	

4.3 Disambiguation of a subset of the BankSearch benchmark dataset

For our experimental studies we chose the pre-classified BankSearch dataset collection [21] consisting of 10,000 web documents classified into 10 equally-sized categories each containing 1,000 web documents. To each category one of four distinct themes, namely Banking and Finance, Programming Languages, Science, and Sport was assigned as shown in Table 3. The dataset is available for download from http://www.pedal.rdg.ac.uk/banksearchdataset/index.htm.

If the term "rule" occurs in a document of this dataset it is usually used in the sense of the assigned theme. Since these themes match nicely with the possible meanings of the term 'rule' as described above (see Table 2), we decided to use these data set for a more detailed quantitative study of the proposed disambiguation method. Therefore, we consider in the following Synset #0 as correctly classifying documents assigned to the banking theme, Synset #1 for the programming theme, Synset #2 for the science and Synset #3 for the sport theme.

For the evaluation, we selected the subset of documents that contain the word "rule". The obtained 369 documents were categorized using the pure Sense Folder classification approach as described in Sect. 3.3 and the clustering approach as described in Sect. 3.4. In both cases we used the revised Synset for "rule" as described in Sect. 4.2 (see Table 2). In Tables 4 and 5 a confusion matrix for the computed Sense Folders and clusters with the a-priori known themes of the Bank Search corpus are given. The "-1" class entries in Table 4 mark documents that have not been classified to any given Sense Folder (for the reasons see the discussion in Sect. 4.1).

Comparing the classification results of Tables 4 and 5 we can see that the clustering approach strongly improved the overall classification performance. While the pure Sense Folder based approach correctly classified only 154 documents (which is a classification rate of correctly classified documents of only 42%) the clustering process was able to assign 258 documents to the correct class, which is an overall performance of approximately 70%. The improvement is especially remarkable for the bank theme, were all documents assigned to the bank theme have been merged correctly in a single cluster. However, also the assignment to all other classes was improvement: At least 8 documents in the bank meaning of the corpus, 19 in Programming, 27 in Science and 5 in Sports have been reclassified to the correct class after the clustering process.

In Table 6 the classification based on the Sense Folder is compared with the class assignments of the clustering method. The elements in the main diagonal (neglecting the column "-1") represent the documents that have been assigned to the same class by the pure Sense Folder and the clustering approach. The other elements of the matrix show the number of documents that have been reassigned to another class. We can see that the majority of documents (244) stayed in the same class. Thus the clustering process seems to be reasonably stable. However, we can also see that all 6 documents that could not be assigned to any class by the Sense Folder approach have been assigned to the wrong cluster (see Table 4 for the correct theme assignments).

Table 3 BankSearch Corpus categories

Dataset Character	Dataset Category	Associated Theme
A	Commercial Banks	Banking & Finance
B	Building Societies	Banking & Finance
C	Insurance Agencies	Banking & Finance
D	Java	Programming Languages
E	C / C++	Programming Languages
F	Visual Basic	Programming Languages
G	Astronomy	Science
H	Biology	Science
I	Soccer	Sport
J	Motor Sport	Sport
X	Sport	Sport

Table 4. Confusion matrix of assignments to Sense Folders and BankSearch themes

	Corpus Theme				
Sense Fold.	Bank	Prog	Science	Sport	Total
-1			3	3	6
0	21	7	11	27	66
1	4	41	19	58	122
2	0	19	40	23	82
3	4	18	19	52	93
Total	29	85	92	163	369

Table 5. Confusion matrix of assignments to clusters and BankSearch themes

	Corpus Theme				
Cluster	Bank	Prog	Science	Sport	Total
0	29	3	8	57	97
1	0	60	3	44	107
2	0	10	67	5	82
3	0	12	14	57	83
Total	29	85	92	163	369

Table 6. Re-classification matrix of Sense Folders and Clusters

	Sense Folder					
Cluster	-1	0	1	2	3	Total
0	6	45	20	10	16	97
1	0	5	84	12	6	107
2	0	7	17	51	7	82
3	0	9	1	9	64	83
Total	6	66	122	82	93	369

5 Conclusions

In this paper we have presented a retrieval system that uses ontologies and a clustering process in order to provide disambiguating information with respect to the given search terms for the documents of a result set. The system assigns subsets of the search results to disambiguating classes ("Sense Folders") that are defined by the employed ontology. The clustering process is used in order to improve the categorization of the documents compared to the classification using only prototype vectors constructed based on the ontology as studied in prior work [6]. So far, the results of our approach are very promising. However, even though we studied the performance using different data sets, a more detailed analysis of the approach using semantically annotated documents is still necessary.

Additionally, we discussed the problem of a “good structured” ontology which is fundamental in order to avoid disambiguation problems with too fine grained ontologies or missing instances like, e.g., the “Range Rover” problem in the context of “range” that can not be solved if the ontology does not include instances like this.

Our approach is query and not collection oriented. The long-term goal is to develop an upgradeable information retrieval system using different ontologies and merging different search procedures.

References

1. Agirre, E., Rigau. G. 1996. Word sense disambiguation using conceptual density. In Proceedings of COLING'96, Copenhagen, Danmark (1996) 16-22
2. Bentivogli, L., Pianta, E., Girardi, C., MultiWordNet: developing an aligned multilingual database, Proceedings of the First International Conference on Global WordNet, Mysore, India (2002)
3. Berners-Lee,T.: Semantic Web Road Map, http://www.w3.org/DesignIssues/Semantic.html (1998)
4. Brusilovsky, P., Methods and techniques of adaptive hypermedia, In: User Modeling and User Adapted Interaction, 6(2-3) (1996).
5. Ciravegna, F., Magnini, B., Pianta, E., Strapparava, C., Multilingual Lexical Knowledge Bases: Applied WordNet Prospects, International Workshop on The Future of the Dictionary, Grenoble (1994).
6. De Luca, E. W., Nürnberger, A., Ontology-Based Semantic Online Classification of Documents: Supporting Users in Searching the Web, In: Proc. of the European Symposium on Intelligent Technologies (EUNITE 2004), Aachen (2004).
7. Ding, Y., van Rijsbergen, C. J., Ounis, I., Jose, J., Report on ACM SIGIR Workshop on ’Semantic Web’ SWIR 2003. Toronto, Canada (2003).
8. Fellbaum, C., WordNet, an electronical lexical database, Cambridge, MA: MIT Press (1998).
9. Fensel, D., van Harmelen, F., Klein, M., Akkermans, A., et al., On-To-Knowledge: Ontology-based Tools for Knowledge Management, Free University Amsterdam, The Netherlands (2000).
10. Guarino, N. and Welty, C. An Overview of OntoClean in S. Staab and R. Studer (eds.), Handbook on Ontologies, Springer Verlag (2004)
11. Gangemi A., Guarino N., Oltramari A., Conceptual analysis of lexical taxonomies: the case of WordNet top-level, Formal Ontology in Information Systems archive, Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001 Ogunquit, Maine, USA (2001) 285 – 296.
12. Gonzalo, J., Verdejo, F., Chugur, I., and Cigarrán, J., Indexing with WordNet synsets can improve text retrieval, in Proc. of the COLING/ACL '98 Workshop on Usage of WordNet for NLP (1998).
13. Klose, A., Nürnberger, A., Kruse, R., Hartmann, G.K., and Richards, M., Interactive Text Retrieval Based on Document Similarities, Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy, 25(8) (2000) 649-654
14. Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A., Comparing Ontology-Based and Corpus-Based Domain Annotation in WordNet, Proceedings of First International WordNet Conference, Mysore, India (2002) 146-154.

15. Motta, E., Buckingham Shum, S. and Domingue, J., Ontology-Driven Document Enrichment: Principles and Case Studies. In Proc. KAW'99: 12th Banff Knowledge Acquisition Workshop, Banff, Canada (1999).
16. Oltramari A., Gangemi A., Guarino N. and Masolo C., Restructuring WordNet's Top-Level: The OntoClean approach, The Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, Spain (2002).
17. Pianta,E., Bentivoglio, L., Girardi, C., MultiWordNet: Developing an Aligned Multilingual Database. In Proceedings of the First International Conference on Global WordNet, Mysore, India, (2002) 293-302.
18. Porter, M., An algorithm for suffix stripping, Program (1980) 130-137.
19. Salton, G., Wong, A., and Yang, C.S., A Vector Space Model for Automatic Indexing, Communications of the ACM, 18 (1975) 613-620.
20. Salton, G., Buckley, C., Term Weighting Approaches in Automatic Text Retrieval, Information Processing & Management, 24(5) (1988) 513-523.
21. Sinka, M.P., Corne, D.W. A large benchmark dataset for web document clustering, in Abraham, A., Ruiz-del-Solar, J., Koeppen, M. (eds.), Soft Computing Systems: Design, Management and Applications, Volume 87 of Frontiers in Artificial Intelligence and Applications (2002) 881-890.
22. Vossen, P., Introduction to EuroWordNet, In: Nancy Ide, N., Greenstein, D. and Vossen, P. (eds), Special Issue on EuroWordNet. Computers and the Humanities, 32(2-3) (1998) 73-89.
23. Wordnet homepage, http://www.cogsci.princeton.edu/~wn/ („Five Papers on Wordnet“) (2004)