Available online at www.sciencedirect.com
ScienceDirect
Web Semantics: Science, Services and Agents on the World Wide Web 5 (2007) 195–203
JOURNAL OF Web Semantics
www.elsevier.com/locate/websem

# Patterns of semantic relations to improve image content search

Laura Hollink [a,*], Guus Schreiber [a], Bob Wielinga [b]

[a] *Section Business Informatics, Free University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands*
[b] *Human Computer Studies Laboratory, University of Amsterdam, Kruislaan 419, 1098 VA Amsterdam, The Netherlands*

Received 24 January 2007; accepted 21 May 2007
Available online 24 May 2007

**Abstract**

This paper reports on a study to explore how semantic relations can be used to expand a query for objects in an image. The study is part of a project with the overall objective to provide semantic annotation and search facilities for a virtual collection of art resources. In this study we used semantic relations from WordNet for 15 image content queries. The results show that, next to the hyponym/hypernym relation, the meronym/holonym (part-of) relation is particularly useful in query expansion. We identified a number of relation patterns that improve recall without jeopardising precision.


*Keywords:* Semantic annotation; WordNet; Query expansion; Image retrieval

## 1. Introduction

The advance of the semantic web enables more and more sophisticated information retrieval. Using ontologies or thesauri, we can now match documents to queries based on semantic similarity, even if there is no textual match between the query and the annotation. Tools have emerged that demonstrate this type of semantic annotation and search. The E-Culture demonstrator, the winner of the Semantic Web Challenge [21], uses several existing vocabularies, such as WordNet [9] and the Art and Architecture Thesaurus (AAT),[1] for annotation and search of a heterogeneous collection of visual resources. Hence, a query for 'flower' will not only return documents about flowers, but also documents annotated with 'roses' since there is a subclass relation between rose and flower. A search for cubist paintings will return paintings annotated with Picasso as the creator, since there are links between creators and styles. A query for 'Venus' will return paintings depicting Aphrodite, if the ontology in the background contains an equivalence relation between the two. All these examples use semantic relations between concepts to improve search results.

While the technology to find such semantically related documents is available, the problem remains that not all *related* documents are *relevant* documents. This becomes apparent from the following examples: a query for pictures of a car will not be satisfied by images of the brakes, even though the brakes are part of the car; a query for 'finish' will not be satisfied with documents annotated with 'start', even though the two are directly related (with an antonym relation from WordNet). This problem becomes bigger when large ontologies or groups of interlinked ontologies are used, which is a realistic and desirable scenario on the semantic web. With large ontologies like WordNet or the AAT, that contain over 15 types of relations, simply returning all documents that are in some way related to the query is no longer an option. Therefore, we recognise a need to investigate which types of semantic relations between a query and a document are likely to improve search results. Moreover, there is a need to study the effect of combinations of relations.

In this paper we address this question in an empirical manner. We focus on semantic relations in WordNet, since this well-known and widely used resource provides a wide variety of relations. We use these relations to find documents that would not be found by just the initial query. A query for Eating, for example, could result in paintings annotated with `banquet`, since in WordNet `wn:banquet` (or feast) is `wn:derivationally_related_to wn:feasting`, which is a `wn:hyponym` of `wn:eating`. Intuitively, the more relations we use to expand the query, the higher recall will be.

* Corresponding author. Tel.: +31 20 5987740; fax: +31 20 5987728.
*E-mail addresses:* hollink@cs.vu.nl (L. Hollink), schreiber@cs.vu.nl (A.Th. Schreiber), wielinga@science.uva.nl (B. Wielinga).
[1] http://www.getty.edu/research/conducting_research/vocabularies/aat/.

1570-8268/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2007.05.002

On the other hand, if too many relations are used, precision will be low. The aim of the present work is to identify which relations give the best balance between recall and precision. We will look into the effect of combinations of relations and the optimal number of nodes between a query concept and an annotation concept.

In an experimental setting, we query a collection of annotated paintings using not only our initial query concepts, but also closely related concepts. In order to discover which relations lead to the best search results, we pose the queries using different types of relations and examine the results. The collection of paintings is a subset of the Artchive collection [12], annotated with the E-Culture web demonstrator. The annotations describe objects that are depicted in the paintings, such as 'man', 'rose' or 'castle'. All annotations correspond to concepts from WordNet.

In Section 2 we provide background information about the E-Culture project and its annotation demonstrator, and give a brief overview of WordNet terminology. Section 3 contains related work on query expansion and semantic annotation. Section 4 contains the design of the experiment and Section 5 the experimental results. We conclude in Section 6, where we reflect on our results and discuss possible generalisations of the findings to other ontologies and collections.

## 2. Background

### 2.1. The E-Culture project

The main objective of the E-Culture project is to employ novel semantic web and presentation techniques to provide better indexing and search mechanisms for the knowledge-rich domain of cultural heritage [21]. All annotations used in the present study were created with the web demonstrator of the E-Culture project. In this section we will elaborate on the annotation features of the E-Culture demonstrator: the vocabularies and the metadata schema.

At the time of writing, the demonstrator uses four vocabularies for annotation and search: the AAT, the Union List of Artist Names (ULAN),[2] the Thesaurus of Geographical Names (TGN)[3] and WordNet 2.0. All were translated from their native format into RDF/OWL.

The metadata schema of the demonstrator is based on Dublin Core [8]. For content annotations, however, Dublin Core is insufficient since it provides only two elements to describe the content of an image. Therefore, we specialised the Dublin Core element `dc:subject` with subproperties to provide more structure in the content descriptions. In the present study we use one of these subproperties, namely `ec:object`. This property describes annotations of objects depicted in the image, such as 'face', 'potato', 'church', but also 'Seine', 'Mme Matisse' or 'Van Gogh'. All annotations used in this study were made using concepts from WordNet.

### 2.2. WordNet

WordNet is a lexical database of the English language. It contains 155,327 English words, namely nouns, verbs, adjectives and adverbs. Many of these words are polysemous, which means that one word has multiple meanings or *senses*. The word `tree`, for example, has three word-senses: tree#1 (woody plant), tree#2 (figure) and tree#3 (English actor). WordNet distinguishes 207,016 word-senses.

Word-senses are grouped into synonym sets (synsets) based on their meaning and use in natural language. Each synset represents one distinct concept. An example of a synset is {cliff#1, drop#4, drop-off#2}, described as "a steep high face of rock". Semantic relations and lexical relations exist between word-senses and between synsets. For the purpose of this paper we will not go into details of all these relations, but rather explain the most common ones. The main hierarchy in WordNet is built on hypernym/hyponym relations between synsets, which are similar to superclass/subclass relations. Other frequent relations are meronym and holonym relations, which denote part-of and whole-of relations, respectively.

WordNet is freely available from the Princeton website.[4] In addition, W3C has released a RDF/OWL representation of WordNet.[5] For easy integration with our annotation interface, and for easy querying with semantic web tools such as SeRQL in Sesame, we use this RDF/OWL version. We treat the RDF/OWL version of WordNet as if it were an ontology, exploiting its size, widespread use and large number of relationships.

## 3. Related work

'Query expansion' is a term from the Information Retrieval community, where it is used as a term for adding related words to a query in order to increase the number of returned documents and with that increase recall. The use of WordNet for expansion of natural language queries for text retrieval has been studied extensively (e.g. [26]). A prerequisite for this type of query expansion is that the correct WordNet word-sense has to be assigned to words in the query. This process is called word-sense disambiguation (WSD).

Voorhees [25] demonstrated that the success of query expansion depends on the length of queries and on the selection of the right synsets. She manually and automatically selected query synsets and expanded these with directly related synsets. In her study she showed that when query synsets were manually selected, recall improved for short queries, but not for longer queries. When query synsets were automatically selected, query expansion did not improve the results at all. Gonzalo et al. [10] measured the sensitivity of retrieval performance to disambiguation errors. They manually indexed both queries and documents with WordNet synsets, deliberately introducing errors. They found that indexing with synsets improved search substantially if the word-sense disambiguation error was less than 10%. A dis-

[2] http://www.getty.edu/research/conducting_research/vocabularies/ulan/.
[3] http://www.getty.edu/research/conducting_research/vocabularies/tgn/.
[4] http://wordnet.princeton.edu/man/wnstats.7WNon 13 December 2005.
[5] http://www.w3.org/TR/wordnet-rdf/.

ambiguation error of more than 30% produced no improvement over using just the original terms in the queries and documents.

Moldovan and Mihalcea [18] developed a method for WSD with 87% accuracy for nouns, which is within the 30% error margin. They expanded short queries with words that belong to the same WordNet synset. Expansion led to an increase in precision for queries provided by the sixth Text Retrieval Conference (TREC), but there was no increase for queries posed by users of internet search engines. Smeaton and Quigley [24] used expansion techniques on image captions. They manually disambiguated words from both queries and captions, and added WordNet synonyms to each word. Retrieval based on these expanded queries and documents gave better results than retrieval based on just the original words.

Although most expansion techniques rely on WordNet synonyms, also hyponyms, hypernyms and words in the glosses have been used. Liu et al. [17], for example, expanded queries with synonyms, hyponyms and glosses and found that this improved results over non-expanded queries. They do not report on the accuracy of their WSD method. Buscaldi et al. [5] expanded geographical terms in queries of GeoCLEF[6] with WordNet synonyms and meronyms. Only meronyms that contained the word 'capital' in the gloss were used. Although their GeoCLEF results were not promising, they pointed out that this type of meronym expansion is most helpful when the geographical names represent political entities.

Few studies compare the effect of different types of relations. Navigli and Velardi [19] compared retrieval results of original queries to results of synset queries and to results of three types of expanded queries: (1) expansion with hyponyms, (2) expansion with synsets of disambiguated gloss words and (3) with plain words from the glosses. They posed 24 queries provided by TREC 2001 to Google. Expansion with plain words from the glosses gave the best results (23% increase over original queries), while the other methods only showed an increase of 1–3% over original queries. They do not report on the accuracy of their WSD method and the effect of this on the results. Sim [22] retrieved URLs, were a URL containing the exact query word is considered most relevant, followed by a URL with a synonym, a hyponym and finally a hypernym. They found that the optimal weight for each query expansion type is 1.0, 0.8, 0.6 and 0.4 for exact words, synonyms, hyponyms and hypernyms, respectively. None of these papers, however, report on the effect of *combinations* of WordNet relations on the results of expanded queries.

The consensus seems to be that WordNet relations improve search only if the correct synsets are used in queries and documents. With the semantic web, a number of retrieval systems have emerged that make this condition a realistic one; they facilitate annotation and search with WordNet synsets or with concepts from other ontologies. Since WSD does not affect retrieval results of these systems, they enable us to take a more detailed look at the effect of different types of relations on retrieval results. Moreover, the semantic web makes it *necessary* to look into these effects since the number of relations in the semantic web is too large to simply return all related documents. We should note that the WSD problem is not solved by semantic annotation and search systems. Rather, it is circumvented by asking a user to express his or her information needs in terms of ontological concepts instead of in natural language.

The E-Culture web demonstrator that was discussed in Section 2.1 is an example of a semantic annotation and search application. Another well-known example is MuseumFinland [15]. This web-based system integrates collections of several Finish museums by translating the existing annotations to concepts from a number of ontologies. The collections can be searched in a multi-faceted browsing interface or with keywords. Alternatively, users are able to search the collection using a multi-facetted thesaurus browser [14]. Sinclair et al. [23] have been working on a portal from which collections of cultural heritage institutions can be searched and annotated with concepts from ontologies. The CIDOC CRM [7] is used as a common framework to integrate the different metadata schemas used by the institutions. Bloehdorn et al. [3] annotate images with a domain ontology, which is linked to a core ontology (DOLCE) and a visual ontology (Mpeg-7). Other examples of semantic annotation and search tools are the Semantic Markup Tool of Kettler et al. [16] and the annotation tool for NASA images of Halaschek-Wiener et al. [11].

Many systems use hyponym, subclass or narrower term (NT) relations to expand queries. Although some systems use more than one type of relation – in MuseumFinland meronyms are used as well as hyponyms – none of them report on the added value of different types of relations for search results, nor on the effect of combinations of relations.

## 4. Experimental setup

In order to find out which (combinations of) relations lead to improvements in search results, we queried a collection of Artchive paintings annotated with WordNet synsets. A total of 202 paintings by 25 painters were annotated by 12 members of the E-Culture project. The annotators were given a set of guidelines to ensure a uniform view on content annotation.[7] The annotators were moderately familiar with the vocabulary (WordNet) and were not aware of the research questions to be answered in the present experiment. The resulting annotations and the RDF/OWL version of WordNet were stored in a Sesame repository and queried with SeRQL [4].

Fifteen query concepts were chosen by looking at objects depicted in paintings in the Artchive collection that were not annotated nor used in the experiment. The query concepts were chosen to be on Rosch's basic level [20]. Concepts at the basic level maximise the number of attributes shared by instances of that concept and minimise the number of attributes shared with other concepts. Apple is a good example of a basic level concept; all apples share a large set of features and are easily distinguishable from other concepts such as bananas. The super-

[6] http://ir.shef.ac.uk/geoclef/.

[7] http://www.cs.vu.nl/~laurah/ECultureGuidelines.pdf.

Table 1
Precision and recall of queries over query types

| Query | Total relevant | Retrieved | | | Correct hits | | | Precision | | | Recall | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ext | Hyp | All | Ext | Hyp | All | Ext | Hyp | All | Ext | Hyp | All |
| Mountain | 15 | 6 | 6 | 30 | 5 | 5 | 10 | 0.83 | 0.83 | 0.33 | 0.33 | 0.33 | 0.66 |
| Window | 49 | 2 | 2 | 64 | 2 | 2 | 28 | 1.00 | 1.00 | 0.44 | 0.04 | 0.04 | 0.57 |
| Cloud | 53 | 2 | 4 | 40 | 1 | 3 | 21 | 0.50 | 0.75 | 0.53 | 0.02 | 0.06 | 0.40 |
| Hand | 56 | 3 | 3 | 62 | 3 | 3 | 31 | 1.00 | 1.00 | 0.50 | 0.05 | 0.05 | 0.55 |
| Male_child | 4 | 1 | 1 | 66 | 1 | 1 | 2 | 1.00 | 1.00 | 0.03 | 0.25 | 0.25 | 0.50 |
| Guitar | 4 | 2 | 2 | 19 | 1 | 1 | 3 | 0.50 | 0.50 | 0.16 | 0.25 | 0.25 | 0.75 |
| Horse | 7 | 1 | 1 | 43 | 1 | 1 | 2 | 1.00 | 1.00 | 0.05 | 0.14 | 0.14 | 0.29 |
| Chair | 12 | 5 | 5 | 35 | 5 | 5 | 7 | 1.00 | 1.00 | 0.20 | 0.42 | 0.42 | 0.58 |
| Woman | 59 | 15 | 20 | 52 | 13 | 17 | 38 | 0.87 | 0.85 | 0.58 | 0.22 | 0.29 | 0.64 |
| Apple | 6 | 2 | 2 | 28 | 2 | 2 | 5 | 1.00 | 1.00 | 0.18 | 0.33 | 0.33 | 0.83 |
| Bottle | 4 | 2 | 2 | 50 | 1 | 1 | 1 | 0.50 | 0.50 | 0.02 | 0.25 | 0.25 | 0.25 |
| House | 37 | 7 | 10 | 53 | 7 | 10 | 22 | 1.00 | 1.00 | 0.42 | 0.19 | 0.27 | 0.59 |
| River | 15 | 6 | 8 | 93 | 5 | 7 | 11 | 0.83 | 0.88 | 0.12 | 0.33 | 0.47 | 0.73 |
| Tree | 49 | 13 | 17 | 30 | 12 | 15 | 17 | 0.92 | 0.88 | 0.57 | 0.24 | 0.31 | 0.35 |
| *Trunk* | *49* | *0* | *0* | *32* | *0* | *0* | *18* | – | – | *0.56* | *0.00* | *0.00* | *0.37* |
| Mean | 26.43 | 4.79 | 5.93 | 48.50 | 4.21 | 5.21 | 14.41 | 0.85 | 0.87 | 0.29 | 0.22 | 0.25 | 0.55 |

class Fruit is more general than the basic level as its instances show large variations. 'Granny Smith' is more specific than the basic level as these apples share many features with other types of apples. It was shown that humans prefer the basic level when verifying if on object belongs to a category, when naming objects and when learning a language [1]. From these findings we hypothesise that the basic level is a natural level for people to query on and therefore a realistic criterion for our set of query concepts.

One query concept, namely Tree, is more general then the basic level. In flora and fauna, the basic level is usually on the level of 'genus', which for trees would have been oak or chestnut. The annotators, however, were not able to distinguish a chestnut from an oak, especially in paintings. This justifies the use of the more general query concept Tree. The 15 query concepts are listed in Table 1. None of the queries were directly related to each other, although some were related through one or more intermediate nodes. Window and House are both related to `wn:building`; Hand, Male_child and Woman are all related to `wn:person`. Each query was posed in three ways:

- *Exact-queries*: only paintings that are annotated with the query concept are returned.
- *Hyponym-queries*: paintings that are annotated with the query concept and paintings annotated with a concept that is related to the query concept through hyponym relations are returned. Up to four intermediate nodes are allowed.
- *All-relations-queries*: paintings that are annotated with the query concept and paintings that are annotated with a concept that is in any way related to the query concept are returned. Up to four intermediate nodes are allowed.

Recall and precision of each query was measured by comparing the results to a golden standard of matching paintings for that query concept. To come to a golden standard, all paintings were judged by two raters. Cohen's Kappa ($\kappa$) was used to measure correspondence between raters. The mean $\kappa$ of all query concepts was 0.68, which is acceptable [6].

## 5. Results

Table 1 shows the number of relevant paintings in the collection (total relevant), the number of retrieved paintings (retrieved), the number of correctly retrieved paintings (correct hits), recall and precision of each query in each condition: exact-queries (Ext), hyponym-queries (Hyp) and all-relations-queries (All). Recall appears to be low for all query types. This is due to the fact that the raters were advised to make the golden standard strict; when a query concept was visible in an image, no matter how small or insignificant, the image was counted as a hit. The annotators, on the other hand, only annotated objects that were clearly visible or important in the image. This frequently led to situations in which raters considered a painting relevant because it depicted an object matching a query concept, but annotators did not annotate the object because it was not important. A painting depicting, for example, an apple and a bottle, could be annotated with just apple, but counted as a correct hit for both apple and bottle. This had a negative effect on recall. Similarly, it might have had a positive effect on precision. Therefore, the recall and precision values of each query type can only be understood in relation to the recall and precision of the other query types.

One of the 15 query concepts, Trunk, was left out of the analysis. It produced no results on exact-queries or hyponym-queries and incorporating it would corrupt the statistical analysis. It was therefore also not used to determine the mean values in Table 1. Nonetheless, Trunk provides a good illustration of the added value of other types of relations than just hyponyms. The fact that Trunk is a `meronym` of `tree` made it possible to return all paintings annotated with `tree` for the query concept Trunk, which lead to high recall (0.37) and precision (0.56).

The three conditions were compared amongst each other with one-way repeated measures analyses of variance (ANOVAs). There was a significant effect of query type on recall ($F(2, 26) = 46.99$, $p < 0.01$). Also, there was a significant effect of query type on precision ($F(2, 26) = 63.8$, $p < 0.01$). Paired $t$-tests showed no significant difference between precision of exact-queries and hyponym-queries. There was a significant difference between precision of exact-queries and all-relations-queries ($p < 0.01$) and between hyponym-queries and all-relations-queries ($p < 0.01$). Paired $t$-tests showed that recall differed between all query types: between exact-queries and all-relations-queries ($p < 0.01$), between hyponym-queries and all-relations-queries ($p < 0.01$) and between exact-queries and hyponym-queries ($p = 0.017$).[8]

The results showed that expansion with `hyponyms` of the query concept increases recall, while maintaining the high precision of exact-queries. The use of other types of relations further increases recall but lowers precision, as was expected. Table 1 shows that for some of the queries in particular, such as Male_child and Horse, precision drops dramatically for all-relations-queries. Closer examination of the results reveals that through a variety of relations, via the intermediate nodes `person` or `body_part`, both Male_child and Horse are connected to a large number of people-related concepts: woman, nude, worker, human_head, torso, etc. These examples confirm the need for a more selective use of relations.

The mean increase in recall of all-relations-queries over hyponym-queries was 0.30 (0.55–0.25). This increase could in part be attributed to the higher number of retrieved images. However, the increase in recall was more than we would expect from the additionally retrieved images only. Suppose that the additional number of retrieved images were randomly taken from the collection, then we would expect an increase in recall of 0.16 according to the following equation[9]:

$$R_{\text{incr.}} = \frac{1}{15}\sum_{i=1}^{15}\frac{(\text{Ret\_All}_i - \text{Ret\_Hyp}_i)(\text{Rel}_i - \text{Hit\_Hyp}_i)}{202 - \text{Ret\_Hyp}_i}\frac{1}{\text{Rel}_i}$$

where $R_{\text{incr.}}$ is the mean expected increase in recall, $\text{Ret\_All}_i$ the number of retrieved images by an all-relations-query for query $i$, $\text{Ret\_Hyp}_i$ the number of retrieved images by a hypernym-query for query $i$, $\text{Rel}_i$ the number of relevant images in the collection for query $i$, $\text{Hit\_Hyp}_i$ the number of hits of a hyponym-query for query $i$ and 202 is the total number of paintings in the collection. Comparing the increase in recall in our experiment to the expected increase in recall based on additionally retrieved images only, we found the experimental values to be significantly higher ($p < 0.01$).

Examining the results of the hyponym- and all-relations-queries, we found that patterns containing four intermediate nodes between query and annotation (which was the maximum in our experiment) were not beneficial to the results: those patterns led to 231 incorrectly retrieved images and only 25 hits. For example, Monet's 'The Thames below Westminster' was incorrectly returned for the query concept Mountain, since it was annotated with `Thames`, which is a `meronym of - England - holonym of - Pennines - hyponym of - hills - hyponym of - natural_elevation - hypernym of - mountain`.

All-relations-queries correctly retrieved 143 paintings that were not found with hyponym queries. The additional hits were caused by 21 distinct patterns of relations (excluding patterns with more than four intermediate nodes). Transitivity of hypernym, hyponym, meronym and holonym relations was assumed to come to the 23 patterns, so `hypernymOf - hyponymOf` and `hypernymOf - hypernymOf - hyponymOf` were counted as the same pattern. We interpreted the WordNet relations `memberHolonym`, `substanceHolonym` and `partHolonym` as one type: `Holonym`. The same was done for different types of `Meronym` relations. Over 90% of the `Meronyms` and `Holonyms` were `partMeronyms` and `partHolonyms`.

The five patterns that led to the most additional hits are depicted in Fig. 1. Pattern 1 returned annotations that are more general than the query concept, such as a still life annotated with Fruit for a query for Apple. The second pattern is called 'siblings'. It linked, for example, a query for Mountain to a painting annotated with Hill, since both are children of `natural_elevation`. Pattern 3 uses part-of relations. It retrieved paintings of Buildings for a query for Window. Pattern 4 combines two types of relations: `hyponym` and `holonym`. An example of a painting that was retrieved by this pattern is 'Wheat Field' by Van Gogh. It contains a `house` which is a `hyponym of - building - holonym of` our query concept Window. Pattern 5 was caused solely by the query concept Hand since WordNet contains the following facts: `person - holonym of - body - meronym of - human - holomym of - hand`. This caused all paintings of people to

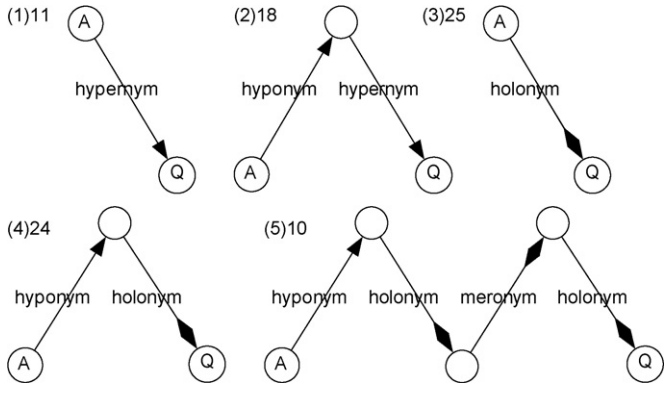


Fig. 1. Patterns of relations that contributed most to recall and the number of correct hits they produced. Note that the black diamond symbol is used to denote both holonym and meronym relations.

[8] In the case of three $t$-tests with d.f.=13 and $\alpha = 0.05$, Bonferoni adjustment calls for a significance level $p$ of at most 0.017. None of our $p$-values exceeded this level.

[9] This equation is derived from the equation $\mathbb{E}x = nm/r$. This problem is also known as the "urn problem", since it asks for the expected number of white balls ($\mathbb{E}x$) out of $n$ balls that are drawn from an urn, containing $m$ white balls and $r - m$ red balls.

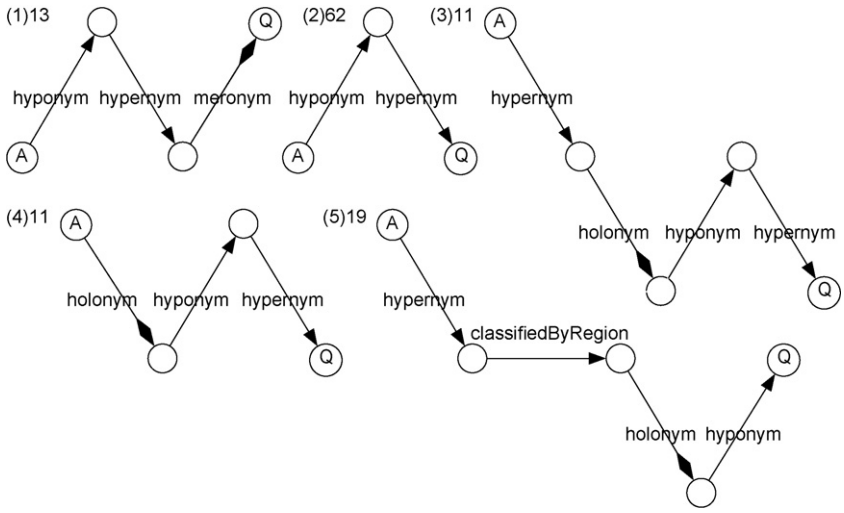


Fig. 2. Patterns of relations that caused low precision and the number of incorrectly retrieved paintings.

be returned for the query Hand. As this structure is present for all body parts, we do not consider this an outlier.

All five successful patterns involve solely `hypernym`, `hyponym`, `holonym` and `meronym` relations. Other types of relations occurring in various patterns led to few hits while resulting in a considerable number of incorrectly retrieved images. Examples are patterns involving `antonym` (5 incorrect, no hits), `inSynset` (7 incorrect, no hits), `classifiedByRegion` (60 incorrect, 1 hit) and `classifiedByTopic` (17 incorrect, 3 hits).[10] Note that the relations `inSynset` and `containsWordSense` are not between two `synsets`, but between `words` and `synsets` or `words` and `word-senses`, respectively. Relations involving `words` or `word-senses` occurred because we did not require intermediate nodes to be `synsets`. However, these relations were rare and did not lead to any hits. `ClassifiedByTopic` was useful only for the query concept River, since it links `wn:river` to `wn:body_of_water`.

Fig. 2 shows the five patterns that lead to the highest number of incorrectly retrieved images. Pattern 5, for example, incorrectly returned 'The Empire of Lights' by Magritte for the query concept River, because the painting contains a house and WordNet has the following statements: `house - hypernym of - maisonette - classified by region - France - holonym of - Loire - hyponym of - river`.

Comparison of Figs. 1 and 2 shows that the pattern `hyponym–hypernym`, also called 'siblings', returns many hits, but even more incorrect images. Siblings are therefore not advantageous for retrieval. Not only siblings, but all other combinations of `hypernym` with another property (e.g. `meronym` or `holonym`) appear disadvantageous. Patterns that combined `hypernym` with another property led to 154 incorrect images and only 28 hits. `Hypernym` alone did give good results. Pattern 1 in Fig. 1 summarises `hypernym` relations with zero or one intermediate node. Longer chains of `hypernym` relations did not occur in our experiment.

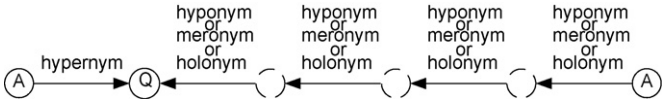


Fig. 3. Proposed query for expansion. Q is the query concept and A represents the annotation concept. Optional intermediate nodes are dashed.

Concluding, it appears that for optimal retrieval results the relation between query concept and annotation concept should be a `hypernym` relation with up to one intermediate node, or any combination of `hyponym`, `meronym` and `holonym` with up to three intermediate nodes. We propose the pattern in Fig. 3 to expand queries with.

We expanded the 15 query concepts with the proposed pattern. Table 2 shows that the proposed query results in a recall of 42% and precision of 64%. The performed *t*-tests showed a sig-

Table 2
Precision and recall of the proposed query

| Query | Total relevant | Retrieved | Correct hits | Precision | Recall |
|---|---|---|---|---|---|
| Mountain | 15 | 8 | 6 | 0.75 | 0.40 |
| Window | 49 | 40 | 24 | 0.60 | 0.49 |
| Cloud | 53 | 22 | 15 | 0.68 | 0.28 |
| Hand | 56 | 38 | 19 | 0.50 | 0.34 |
| Male_child | 4 | 5 | 1 | 0.20 | 0.25 |
| Guitar | 4 | 9 | 3 | 0.33 | 0.75 |
| Horse | 7 | 1 | 1 | 1.00 | 0.14 |
| Chair | 12 | 12 | 6 | 0.50 | 0.50 |
| Woman | 59 | 28 | 23 | 0.82 | 0.39 |
| Apple | 6 | 5 | 5 | 1.00 | 0.83 |
| Bottle | 4 | 7 | 1 | 0.14 | 0.25 |
| House | 37 | 18 | 16 | 0.89 | 0.43 |
| River | 15 | 9 | 7 | 0.78 | 0.47 |
| Tree | 49 | 20 | 16 | 0.80 | 0.33 |
| *Trunk* | *49* | *19* | *17* | *0.89* | *0.35* |
| Mean | 26.43 | 15.86 | 10.21 | 0.64 | 0.42 |

[10] We use the WordNet property names as published in van Assem [2]. Explanation of the WordNet terminology can also be found in the WordNet manual on http://wordnet.princeton.edu/man/wngloss.7WN.

Table 3
Precision, recall and $F_1$-measure of the four query types

| Query type | Precision | Recall | $F_1$ |
|---|---|---|---|
| Exact | 0.85 | 0.22 | 0.33 |
| Hyponym | 0.87 | 0.25 | 0.36 |
| All-relations | 0.29 | 0.55 | 0.33 |
| Proposed | 0.64 | 0.42 | 0.46 |

nificant difference between hyponym queries (Table 1) and the proposed query for precision ($p < 0.01$) and recall ($p < 0.01$). This shows that query expansion with the right types of relations can improve recall with almost 70% over expansion with only hyponym relations (from 0.25 to 0.42), while preserving an acceptable level of precision.

Although precision and recall are good measures of retrieval performance, their often opposing values make it hard to interpret the value of a retrieval strategy as a whole. As we have seen in the present experiment, when precision goes up, recall goes down, and vice versa. An indication of the overall performance of a retrieval strategy is the $F_1$-measure, which is the harmonic mean of precision and recall, as in the following equation:

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

The mean $F_1$-scores of exact-queries, hyponym-queries, all-relations queries and the proposed queries were 0.33, 0.36, 0.33 and 0.46, respectively (Table 3). Although a significant increase in $F_1$ of the proposed query over hyponym-queries could not be proven ($t = -1.95$, d.f.$= 13$, $p = 0.07$), the numbers indicate that the proposed query performs better than the other expansion strategies.

## 6. Discussion

Although the pattern was obtained in an empirical manner, the characteristics of the pattern can be explained also from a conceptual point of view. The proposed number of nodes between a query and an annotation is markedly shorter when going up in the hypernym/hyponym hierarchy than when going down. The proposed query pattern in Fig. 3 recommends a direct link between a query and an annotation when going up, while there can be up to three intermediate nodes when going down. Although the exact length of the pattern depends on the specific vocabulary – in our case WordNet – a general rule seems to be that one should be more conservative with expansion by going up in the hierarchy than by going down towards more specific concepts. The depth of the hypernym/hyponym hierarchy varies greatly in WordNet. Most parts of the hierarchy are relatively shallow, but some parts, such as the hierarchies of flora and fauna, are more than 14 levels deep. In our experiment, we tested chains of relations between query concept and annotation concept of up to four intermediate nodes. We found that four intermediate nodes performed worse than up to three intermediate nodes. However, we suspect that in some cases a deeper approach of up to seven or eight intermediate nodes will make a positive difference on retrieval of concepts in deep hierarchies such as plants and animals. In our study, a query for Plant, for example, could not be related to Apple Tree because they are related with more then four intermediate nodes. An alternative strategy that needs additional testing is to allow hyponym relations to have an arbitrary depth. Some databases interpret hyponym relations as a transitive property and pre-compute the complete transitive closure. In those cases, the length of the pattern of hyponym relations will not cause the query to be computationally expensive. The Mia demonstrator described in [13] applies this strategy.

For part-of relations we observe a different pattern. Going up and going down in the hierarchy is equally beneficial. For retrieval of visual resources, a possible explanation is that in many cases part-of relations show an 'inheritance of visibility' that goes both up and down the hierarchy: if the whole is visible, the parts can be visible as well; if a part is visible, the whole can be visible as well. This holds for many examples, such as hand–finger, house–roof or flock–sheep, but not for internal parts like organs nor for portraits in which the head is visible but not the body. In brief, our experiments clearly illustrate the importance of part-of relations for retrieval of visual resources, but the underlying mechanisms are not yet revealed. Future research is necessary to deepen our understanding of when to use part-of relations, and to verify if part-of relations are equally important for text retrieval as they are for image retrieval.

In the present experiment, the gain in recall caused by expansion was more promising than what was found by the text retrieval community (see Section 3). This might be due to the fact that in our application both queries and annotations were short, often consisting of only a few concepts. Voorhees [25] found that the effect of expansion is higher for short queries than for long queries. The same might hold for the length of annotations. This suggests that query expansion is especially fruitful for image retrieval, that typically involves short documents (annotations) and short queries.

The results of the experiment are not specific to the E-Culture demonstrator, since the experiment did not rely on this system, other than for collecting the annotations. Also, we expect that the results can be generalised to other visual domains than the painting domain. The annotations consisted mostly of everyday concepts that occur in numerous other domains, such as news, collections of photographs, movies, etc. We recognise that the specific structure of WordNet, such as the depth of the hierarchy and the frequency of certain types of relations, has influenced the pattern of semantic relations that was the outcome of this study. However, the types of relations that proved most important in this study – hyponym/hypernym and meronym/holonym – occur frequently in a variety of other vocabularies, sometimes explicitly, such as in the Gene Ontology, and often implicitly as Broader/Narrower Term relation such as in the AAT or MeSH.[11] This is an indication that the same pattern might be beneficial for expansion with other vocabularies.

[11] http://www.nlm.nih.gov/mesh/.

## 7. Conclusion

We examined the use of various WordNet relations and concluded on patterns of relations that proved most beneficial for query expansion.

Expanding queries with hyponyms is intuitive and frequently used by search tools. The present study showed that it indeed improves recall while maintaining precision. The results also show that recall of retrieval results can be further improved if other types of relations are used as well. Expansion with a combination of hyponym, holonym and meronym relations improves recall while maintaining an acceptable level of precision. Likewise, expansion with hypernym relations improves search results. However, a combination of hypernyms with other types of relations (e.g. hyponyms or holonyms) is more detrimental to precision than it is beneficial to recall. Expansion with other types of WordNet relations, such as inSynset and classifiedByRegion, appeared to harm the results. We can conclude that semantic annotation and search systems can improve their recall values by expanding query results with not only hyponym relations, but also with part-of relations and hypernym relations.

The results of the present study can also be used to improve ranking of result sets. Images linked to a query concept through a pattern that produces high precision would then appear higher in the result list than images linked through a pattern that causes low precision. Images linked through, for example, the 'all-relations-query' would end up at the bottom of the list. For queries that yield very little results, expansion with patterns that cause low precision might still be advantageous.

## Acknowledgements

We would like to thank Alia Amin, Mark van Assem, Michiel Hildebrand, Zhicheng Huang, Janneke van Kersen, Borys Omelayenko, Jacco van Ossenbruggen, Novan Pavlovich, Ronny Siebes and Jan Wielemaker from the E-Culture team for creating the annotations. Special thanks go to Jan Wielemaker, Ronny Siebes, Jacco van Ossenbruggen, Victor de Boer, Mark van Assem and Alia Amin for their valuable contributions to the demonstrators, the vocabularies and the experiment. Thanks to Alistair Vardy for rating the images to produce the golden standard.

## References

[1] A. Archambault, F. Gosselin, P. Schyns, A natural bias for the basic level? in: Proceedings of the 22nd Annual Conference of the Cognitive Science Society, 2000, pp. 585–590.

[2] M. van Assem, A. Gangemi, A.Th. Schreiber, RDF/OWL representation of WordNet, W3C editor's draft, April 2006, Electronic document, accessed May 2006, available from: http://www.w3.org/2001/sw/BestPractices/WNET/wn-conversion.html.

[3] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab, M.G. Strintzis, Semantic annotation of images and videos for multimedia analysis, in: Proceedings of the Second European Semantic Web Conference, May 29–June 1, 2005, pp. 592–607.

[4] J. Broekstra, A. Kampman, SeRQL: a second generation RDF query language, in: Proceedings of the SWAD-Europe Workshop on Semantic Web Storage and Retrieval, Amsterdam, The Netherlands, November, 2003, pp. 13–14.

[5] D. Buscaldi, P. Rosso, E. Sanchis Arnal, A WordNet-based query expansion method for geographical information retrieval, in: Working notes of the Cross Language Evaluation Forum (CLEF), Vienna, Austria, September, 2005, pp. 939–946.

[6] J. Carletta, Asessing agreement on classification tasks: the kappa statistic Comput. Ling. 22 (2) (1996) 249–254.

[7] M. Doerr, The CIDOC CRM—an ontological approach to semantic interoperability of metadata, AI Magazine 24 (3) (2003) 75–92.

[8] Dublin Core, Dublin Core metadata element set, version 1.1: reference description. Dublin Core Metadata Initiative, 2006, Electronic document, accessed January 2006, available from: http://dublincore.org/documents/dces/.

[9] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, USA, 1998.

[10] J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarran, Indexing with WordNet synsets can improve text retrieval, in: Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP, August, 1998, pp. 38–44.

[11] C. Halaschek-Wiener, A. Schain, J. Golbeck, M. Grove, B. Parsia, J. Hendler, A flexible approach for managing digital images on the semantic web, in: Proceedings of the Fifth International Workshop on Knowledge Markup and Semantic Annotation SemAnnot, November, 2005, pp. 49–58.

[12] M. Harden, Mark harden's artchive, Electronic document, accessed April 2006, available from: http://www.artchive.com/.

[13] L. Hollink, A.Th. Schreiber, J. Wielemaker, B.J. Wielinga, Semantic annotation of image collections, in: Proceedings of the K-Cap 2003 Workshop on Knowledge Markup and Semantic Annotation, October, 2003.

[14] E. Hyvönen, M. Junnila, S. Kettula, E. Mäkelä, S. Saarela, M. Salminen, A. Syreeni, A. Valo, K. Viljanen, Finnish museums on the semantic web: the user's perspective on MuseumFinland, in: Proceedings of Museums on the Web, March–April, 2004, pp. 21–32.

[15] E. Hyvönen, M. Salminen, M. Junnila, S. Kettula, A content creation process for the semantic web, in: Proceedings of the LREC Workshop on Ontologies and Lexical Resources in Distributed Environments, May, 2004.

[16] B. Kettler, J. Starz, W. Miller, P. Haglich, A template-based markup tool for semantic web content, in: Y. Gill, E. Motta, V.R. Benjamins, M.A. Musen (Eds.), Proceedings of the Fourth International Semantic Web Conference, November, 2005, pp. 446–460.

[17] S. Liu, F. Liu, C. Yu, W. Meng, An effective approach to document retrieval via utilizing WordNet and recognizing phrases, in: Proceedings of the 27th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 266–272, ISBN: 1-58113-881-4, http://acm.org/doi:10.1145/1008992.1009039.

[18] D.I. Moldovan, R. Mihalcea, Using WordNet and lexical operators to improve internet searches, IEEE Internet Comput. 4 (1) (2000) 34–43, ISSN: 1089-7801, http://dx.doi.org/doi:10.1109/4236.815847.

[19] R. Navigli, P. Velardi, An analysis of ontology-based query expansion strategies, in: Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia, September, 2003, pp. 42–49.

[20] E. Rosch, Basic objects in natural categories, Cogn. Psychol. 8 (3) (1976) 382–439.

[21] A.Th. Schreiber, A. Amin, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, L. Hollink, Z. Huang, J. van Kersen, M. de Niet, B. Omelayenko, J. van Ossenbruggen, R. Siebes, J. Taekema, J. Wielemaker, B.J. Wielinga, Multimedian e-culture demonstrator, in: The Semantic Web Challenge at the Fifth International Semantic Web Conference, Athens, GA, USA, November, 2006.

[22] K.M. Sim, Toward an ontology-enhanced information filtering agent, ACM SIGMOD Record 33 (1) (2004) 95–100, ISSN: 0163–5808, http://acm.org/doi:10.1145/974121.974138.

[23] P. Sinclair, P. Lewis, K. Martinez, M. Addis, D. Prideaux, D. Fina, G. Da, Bormida, eCHASE: sustainable exploitation of electronic cultural heritage, in: Proceedings of the Second European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, November–December, 2005.

[24] A.F. Smeaton, I. Quigley, Experiments on using semantic distances between words in image caption retrieval, in: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, August, 1996, pp. 174–180.

[25] E. Voorhees, Query expansion using lexical–semantic relations, in: W.B. Croft, C.J. van Rijsbergen (Eds.), Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Springer–Verlag, New York, NY, USA, 1994, pp. 61–69.

[26] E.M. Voorhees, The TREC question answering track, Natural Lang. Eng. 7 (4) (2001) 361–378.