Multimed Tools Appl
DOI 10.1007/s11042-009-0400-y

Thesaurus enrichment for query expansion in audiovisual archives

Laura Hollink · Véronique Malaisé · Guus Schreiber

Abstract It is common practice in audiovisual archives to disclose documents using metadata from a structured vocabulary or thesaurus. Many of these thesauri have limited or no structure. The objective of this paper is to find out whether retrieval of audiovisual resources from a collection indexed with an in-house thesaurus can be improved by enriching the thesaurus structure. We propose a method to add structure to a thesaurus by anchoring it to an external, semantically richer thesaurus. We investigate the added value of this enrichment for retrieval purposes. We first anchor the thesaurus to an external resource, WordNet. From this anchoring we infer relations between pairs of terms in the thesaurus that were previously unrelated. We employ the enriched thesaurus in a retrieval experiment on a TRECVID 2007 dataset. The results are promising: with simple techniques we are able to enrich a thesaurus in such a way that it adds to retrieval performance.

Keywords Thesaurus · Retrieval · Ontology alignment · Multimedia · Query expansion

1 Introduction

The objective of this paper is to investigate whether retrieval of audiovisual documents that are indexed with an in-house thesaurus can be improved by enriching the thesaurus structure. We propose to add structure to a thesaurus by anchoring it to an external, semantically richer thesaurus.

L. Hollink (✉) · V. Malaisé · G. Schreiber
VU University Amsterdam, de Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
e-mail: hollink@cs.vu.nl

V. Malaisé
e-mail: vmalaise@few.vu.nl

G. Schreiber
e-mail: schreiber@cs.vu.nl

Published online: 22 October 2009

Many collections of audiovisual documents are indexed manually using terms from a local thesaurus. The manual indexing process is time-consuming, therefore the tendency is to only use a small set of terms to describe a document. The annotations are usually of high quality. We point out that the opposite can be said about automatic annotation using content-based feature detectors. This approach results in many annotations, but their quality varies.

A low number of annotations per document can lead to low recall of search results. One way to overcome this issue is query expansion, where documents are retrieved not only with the initial query term, but also with closely related terms [29]. In the context of concept-based search, where queries are posed in terms of thesaurus concepts, query expansion depends on a rich thesaurus structure. However, local thesauri are often limited in breadth and depth. In this paper we report on an experiment in which we enrich a local thesaurus and study its added value for retrieval.

The study is performed on a dataset of the Netherlands Institute for Sound and Vision (Sound & Vision). The institute stores over 700,000 hours of Dutch broadcast video, and archives every day the daily broadcast in digital format. It has an in-house Dutch language thesaurus, the GTAA, with limited structure. The GTAA is used by a team of professional catalogers to index the collection. They are instructed to focus on the core topic of the video, and typically use only a small set of GTAA terms to describe each audiovisual document. The indexes are searched by broadcast professionals, who reuse material to create new television programs, and by the general public. Testing our approach on the Sound & Vision collection allows us to demonstrate the benefit of an enriched thesaurus for retrieval on a real-life dataset.

Our approach consists of two steps. First, we anchor the thesaurus to an external resource, the English-American WordNet [5], by searching for related concepts (*synsets* in WordNet) using a syntactic alignment procedure. The alignment is based on a lexical comparison of term descriptions in the two resources, following the approach in [15], and employs a freely accessible online bilingual dictionary to enable the anchoring of thesauri in different languages. Such a mainly lexical alignment approach is bound to be incomplete and at times incorrect. We did not correct the mistakes in the alignment, but rather investigated how we can use the anchoring given its short-comings. Considering the complexity of automatically aligning two resources that differ in language, scope and structure, working with an imperfect alignment is a realistic situation that is in line with the state of the art of ontology alignment [4]. In the second step, based on this anchoring, we enrich the in-house thesaurus by inferring potential new relations between terms within the thesaurus.

To investigate the value of the enriched thesaurus for retrieval purposes, we perform an experiment in which we compare retrieval results achieved with the in-house thesaurus to results obtained with the enriched thesaurus. The experiment is performed on a part of the collection of Sound & Vision that was used in the TRECVID 2007 conference [23]. We use the queries and ground truth of TRECVID. In addition, Sound & Vision provided the metadata of this dataset in the form of manual annotations of the audiovisual documents with GTAA terms.[1] Our

[1]The metadata is made available to education and research institutes through the ACADEMIA project (http://www.academia.nl) for a small license fee.

hypothesis is that anchoring the in-house thesaurus to a rich external source will help retrieval, particularly with respect to recall: the richer semantic structure should lead to more matches. We are interested in finding out how much this approach jeopardizes precision and whether the joint effect can be judged to be positive or negative.

The present paper is an extension of earlier work, presented at the SAMT 2008 conference [8]. We have extended both steps of our approach—anchoring and enrichment—and increased the scale of the retrieval experiment. Regarding the first step, the GTAA has been anchored to WordNet more firmly by a more extensive set of mappings. Our continuous work on the alignment, as well as mappings contributed by the DSSIM team [21], a participant of the Ontology Alignment Evaluation Initiative [3], have led to a set of mappings that is not only larger, but also more diverse. In addition, we performed a manual evaluation of the different types of mappings. We discuss the alignment in Section 2.

In the second step, we used this new anchoring to infer three times as much new relations within the GTAA compared to our earlier work. As in the previous step, the paper was extended with a manual evaluation: we judged the quality of the newly inferred relations, taking into account the different types of mappings that they were based on. This is discussed in Section 4.

We repeated the retrieval experiment on the extended set of inferred relations. The conclusions confirm what was found in our previous work, but the larger number of inferred relations allows us to draw more statistically significant conclusions. Section 5 describes the experimental setup, the TRECVID dataset, and the results of the retrieval experiment. We conclude with a discussion and directions for future work in Section 6.

2 Thesauri

Semantic query expansion depends on a rich thesaurus with many interrelated terms. The first step in our approach is to anchor the weakly structured GTAA thesaurus that is used to index and search the audiovisual collection to the larger, semantically richer WordNet. In this section we present both resources.

2.1 The GTAA thesaurus

The GTAA is a Dutch, faceted thesaurus resulting from the merging of several controlled vocabularies used by audiovisual archives in the Netherlands. Its name is a Dutch acronym for “Common Thesaurus for Audiovisual Archives”. At the Netherlands Institute for Sound & Vision, it is used for manual annotation of the extensive collection of broadcast video.

The GTAA thesaurus contains approximately 160,000 terms, organised in six facets: `Location`, `Person name`, `Name`, `Maker`, `Genre` and `Subject`. `Location` describes either the place(s) where the video was shot, the places mentioned or seen on the screen or the places the video is about. `Person name` is used for people who are either seen or heard in the video, or who are the subject of the program; `Name` has the same function for named organisations, groups, bands, periods, events, etc. `Maker` and `Genre` describe the creators and genre of a TV program. The `Subject`

facet is used to describe what a program is about and what can be seen in the video, and aims to contain terms for all topics that could appear on TV, which makes its scope quite broad.

The focus of the present paper is on the Subject facet, since our main aim is to retrieve video based on what it is about. Although other facets could contribute to this aim, the Subject facet is the only facet with semantic relations between its terms, making it the most suitable facet for our method and experimental setup. It is a typical example of an in-house thesaurus in the cultural heritage field, comparable in size and type of semantic relations to, for example, the Brinkman thesaurus of the Dutch Royal Library and the UNESCO thesaurus. The Subject facet contains 3,878 terms. It is organised according to the semantic relations defined in the ISO-standard 2788 for thesauri [11], namely `Broader Term` (linking a specialized concept to a more general one), its inverse relation `Narrower Term` (linking a general concept to a more specialized one), and `Related Term` (denoting an associative link). The GTAA contains 3,382 broader/narrower relations and 7,323 associative relations between terms in the subject facet. The broader/narrower hierarchy is shallow: 85% of the hierarchy is no more than three levels deep. For integration purposes, we used a version of the GTAA that was converted to SKOS in an earlier effort [1]. SKOS provides a common data model to represent thesauri using RDF and port them to the Semantic Web [20].

2.2 WordNet

WordNet is a lexical database of the English language. It currently contains 155,287 English words: 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 206,941 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. Each synset is accompanied by a 'gloss': a definition and/or some example sentences.

WordNet is freely available from the Princeton website.[2] In addition, W3C has released a RDF/OWL representation of WordNet version 2.0.[3] In this study we use this RDF/OWL version, as it allows us to use Semantic Web tools and standards to query the WordNet database.

[2] http://wordnet.princeton.edu/

[3] http://www.w3.org/TR/wordnet-rdf/

Two Dutch resources exist that are highly comparable to the Princeton WordNet: the Dutch part of EuroWordNet[4] and Cornetto.[5] We did not use these for two reasons. First, they are smaller than the Princeton Wordnet, with 44,015 and 70,371 synsets respectively. Moreover, a link to a well known and much used resource such as Princeton WordNet has the advantage that it opens possibilities for links to other resources that are anchored to WordNet, such as a large part of the Linked Data cloud.[6]

3 [Step 1] Anchoring

Anchoring GTAA to WordNet is non-trivial, since the two resources are in different languages. In this section we present how we approach this problem by using resources freely available on the web.

3.1 Approach

The anchoring process starts with a terminological enrichment phase. We used a Dutch lexical database (Celex) to find alternative forms for the terms in the thesaurus. Because the original preferred terms and non-preferred terms in the GTAA are in plural form, we added singular forms. The set of GTAA terms was further extended by splitting compound terms into separate words (again using Celex) and searching two online dictionaries[7,8] for synonymous forms. The list of possible synonyms was not further processed, but simply taken 'as is' as additional candidate labels for the GTAA terms. All forms, the original ones as well as the newly added ones, were used for the anchoring to WordNet as this increases the possible coverage of the mappings.

Next, we queried an online bilingual dictionary[9] for the Dutch terms, which provided English translations and one-sentence descriptions. Finally, we anchored the—now English—GTAA terms to WordNet. In contrast to many anchoring methods (e.g. [13]), we do not compare the terms from the two thesauri, but measure the lexical overlap of their descriptions. The approach is derived from [15], who disambiguated a word by comparing the lexical overlap of all of its possible definitions with the possible definitions of the words that are its neighbor in the sentence. The approach was later followed by [14]. For the comparison of definitions, the similarity measure can range from the percentage of words that occur in both definitions, which was used in [16], to cosine similarity between vectors of words in the definitions. Much to our surprise, in the present case the one-sentence descriptions of the online dictionary *are* the WordNet glosses for 99% of the words, giving us the luxury to

[4] http://www.illc.uva.nl/EuroWordNet/

[5] http://www2.let.vu.nl/oz/cltl/cornetto/license.html

[6] http://linkeddata.org/

[7] http://www.muiswerk.nl/WRDNBOEK/INHOUD.HTM

[8] http://www.vandale.nl/vandale/opzoeken/woordenboek/

[9] http://lookwayup.com

avoid the choice of a similarity measure. The anchoring process is described in more detail in [17].

GTAA terms that were found to correspond to multiple WordNet synsets were anchored to all those synsets. There were three reasons why we didn't attempt to do sense disambiguation. First, we are aiming for an increased recall so our primary focus is finding correct correspondences rather than avoiding incorrect correspondences. Second, disambiguation of terms with little context (which is the case for the GTAA terms) is difficult. In the future, we intend to take into account the broader terms in the thesaurus for disambiguation purposes. The third and most important reason is that linking to more than one synset is often correct because WordNet makes finer distinctions than the GTAA. For example, WordNet distinguishes four meanings for the GTAA term "chicken", described by the glosses: 'adult female chicken', 'the flesh of a chicken used for food', 'a domestic fowl bred for flesh or eggs' and 'a domesticated gallinaceous bird thought to be descended from the red jungle fowl'. This fine-grained distinction is absent in the GTAA. In a similar fashion, some WordNet synsets are linked to more than one GTAA term. For example, the WordNet synset "Studio" was an anchor for both the GTAA terms "Atelier" and "Studio". The anchoring process described here has previously been applied to anchor other Dutch and English thesauri [17]; it can easily be adapted to other terminological resources.

To further extend our set of mappings, we made use of the output of the Ontology Alignment Evaluation Initiative (OAEI) 2008 Campaign.[10] Miklos Nagy kindly gave us permission to use the mappings from GTAA to WordNet that were made with the DSSIM algorithm [21] in the context of the OAEI Very Large Crosslingual Resources track. The DSSIM mappings link a GTAA term to only one WordNet synset; about half of these mappings overlap with mappings that were found by our method. Taking the union of the two sets, 2173 GTAA terms were anchored to at least one WordNet synset, which is more than half of all terms in the Subject facet. As one GTAA term can be mapped to multiple WordNet synsets, the total number of proposed mappings is much larger: 4482.

3.2 Small-scale evaluation of the anchoring

We took a random sample of 100 mappings from the total set of mappings (exclusing the DSSIM mappings), and evaluated these manually. To get a slightly more fine-grained measure, we scored the mappings on a three point scale of "incorrect" (scoring 0), "partially correct" (scoring 0.5), and "correct" (scoring 1), instead of the usual dichotomous 0/1 scores. "Partially correct" mappings link a term to a more generic notion, a more specific one or a term related in its application domain, such as Ship and Captain. This is especially appropriate for our set of mappings, in which we aimed for a high recall rather than a high precision by including also matches based on synonyms and parts of compound terms. We used *generalized precision* as defined by [12] to summarize the scores, which is calculated as the mean of all scores. A sample of 100 of the DSSIM mappings were evaluated using the same three point scale in the OAEI 2008, and we used the scores of this evaluation.

[10] http://oaei.ontologymatching.org/2008/

Table 1 shows the type of lexical information that was used to find the 4482 mappings. The numbers in the table sum up to more than 4482, as some mappings were found in more than one way. For example, the Dutch GTAA term `Ambassades` was mapped to WordNet `Embassy` based on its singular form and based on a synonym. For each type of lexical information, the table shows the number of mappings that ended up in our sample and the results of the evaluation. Mappings based on the original GTAA preferred terms score well (74%). Also, the singular forms derived from Celex score high (73%). From OAEI 2009 we know that the DSSIM precision scores were in the same range (75%). This strengthens our belief that the quality of the alignment does not depend on the surprisingly large overlap between the WordNet glosses and the one-sentence descriptions in the online dictionary that we used in this case (see Section 3).

Synonyms appear to be an unreliable source of mappings (35%). A possible explanation is that synonyms are only valid in a given context, while the terms in the thesaurus are isolated from textual contexts. The use of synonyms without filtering them based on the meaning of the term under consideration, magnifies the problem of ambiguous terms. A possible direction for future research could be to use the broader, narrower and related terms in the thesaurus as the context of a term, in order to select the relevant synonyms. Alternatively, a manual evaluation and correction of the synonyms would alleviate the problem, and at the same time enrich the GTAA with additional synonymous words.

The worst precision score is observed for the split forms of compound terms (21%). The major source of error is the fact that some complex or composed terms *should not* be split. For example, the Dutch word for potato is *aardappel*—literally earth-apple—is split by our algorithm into *aard* and *appel*, leading to erroneous mappings. One simple heuristic could be to consider the mappings based on split forms only when no mapping is found for the full term itself. Despite the difficulties, we believe the split terms can be a valuable addition to the anchoring process. They generate mappings that could not easily have been found in another way. The GTAA term *kindermishandeling*—child molesting—, for example, was mapped to WordNet *maltreatment* based on our compound splitter. In the absence of an exact mapping, we can see this match as useful for query expansion.

Although we recognize the possibility (and the need) for more evaluation—a larger sample size, including recall scores, etc.—this is outside the scope of the current paper. As a start, we did an informal inspection of a small sample of the 'one-to-many' mappings: the number of synsets that is aligned with a particular GTAA term does not seem to be an indication of the quality of the matches; GTAA terms that are matched to multiple synsets are equally well matched as GTAA terms that are matched to only one synset.

Table 1 Number of mappings, evaluated sample size and precision scores of mappings per type of match

	Pref. terms	Sing. forms	Synonyms	Split compounds	DSSIM
No. of map.	712	1530	689	967	611
Sample size	19	40	20	21	*100*
Precision	74%	73%	35%	21%	*75%*

The DSSIM values are in italics because they were evaluated in the context of OAEI 2008

4 [Step 2] Thesaurus enrichment

Thesaurus enrichment has been studied in many forms, one of the most well-known being the use of Hearst patterns to discover hyponyms (or subclasses) [6]. This approach was later extended by machine learning of the patterns (e.g. [25]) and has been applied to a wide range of use cases. For example, in [16], we have explored applying lexico-syntactic patterns to term definitions to discover semantic relations. Also other types of relations than hyponyms or subclasses have been discovered using Hearst-like patterns. Van Hage [27], for example, learned patterns to discover part-whole relations. In this paper, we explore another direction: instead of using the information implicit in texts or on the web, we aim to use the information explicit in a rich semantic resource. Based on the anchoring of the in-house GTAA thesaurus to the much larger, richer resource WordNet, we infer new relations within the GTAA. Using SeRQL [2] queries we relate pairs of GTAA subject terms that were not previously related. This approach is appealing since it enables us to reuse some of the effort that went into the careful construction of WordNet for the expansion of queries on the Sound & Vision archives.

4.1 Approach

Figure 1 illustrates how a relation between two terms in the GTAA, t_1 and t_2, is inferred from their correspondence to WordNet synsets w_1 and w_2. If t_1 corresponds to w_1 and t_2 corresponds to w_2, and w_1 and w_2 are closely related, we infer a relation between t_1 and t_2. The inferred relation is symmetric, illustrated by the two-way arrow between t_1 and t_2.

Two WordNet synsets w_1 and w_2 are considered to be 'closely related' if they are connected through either a direct (i.e. one-step) relation without any intermediate synsets or an indirect (i.e. two-step or three-step) relation with one or two intermediate synsets. The latter situation is shown in Fig. 1. From all WordNet relations, we used only meronym and hyponym relations, which roughly translate to part-of and subclass relations, and their inverses holonym and hypernym. A previous study demonstrated that other types of WordNet relations do not improve retrieval

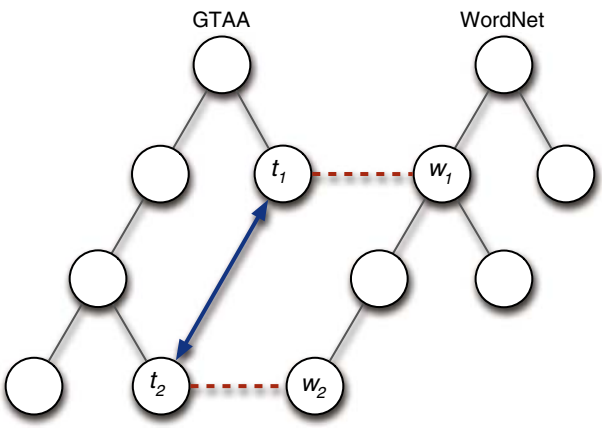

Fig. 1 Using the anchoring to WordNet to infer relations within the GTAA

results when used for query expansion [9]. Both meronym and hyponym can be considered hierarchical relations in a thesaurus. Only sequences of two relations are included in which each has the same direction, since previous research showed that changing direction, especially in the hyponym/hypernym hierarchy, decreases semantic similarity significantly [7, 9]. For example, w_a hypernym of w_b hyponym of w_c is not included.

4.2 Newly inferred relations

A total of 3735 pairs of GTAA terms was newly related: 1206 with one step between WordNet synsets w_1 and w_2, 1378 with 2 steps and 1151 with three steps between w_1 and w_2. 85% of the relations were derived from hyponym relations, 6% from meronym relations, which is a more rare relation in WordNet, and 9% from a combination of hyponym and meronym relations. The number of inferred relations is comparable to the number of existing Broader/Narrower relations in GTAA (3,382).

Inferred relations between pairs of GTAA terms that were already each others `Broader Term`, `Narrower Term` or `Related Term` were not included in the retrieval experiment, nor in the above numbers, since they do not *add* to the structure of the GTAA. 512 relations were discarded for this reason, 304 of which were `Broader/Narrower` and 208 `Related Terms`. On the other hand, inferred relations between GTAA terms that were already each others siblings—e.i. have a common `Broader Term`—were included. The reason is that if we interpret `Broader Term` as a transitive relation (cf. subclass relations), most GTAA terms would be (indirect) siblings of each other because somewhere in the hierarchy of broader terms they have a common parent. Removing all relations between (indirect) siblings would mean discarding possibly interesting relations. Of the total set of inferred relations, only 6% turned out to be between a pair that was already each others direct sibling. This small number is in line with how we inferred the relations from WordNet: only sequences of two WordNet relations are included in which each has the same direction, which is not the case for siblings (see Section 4.1).

An informal inspection of the newly inferred relations quickly leads to a list of examples that we expect to be useful, but also to a number of relations that are incorrect. Table 2 enumerates some good and bad examples. We did not detect a difference in quality between relations inferred from hyponyms and those inferred from meronyms. In the next section we will proceed to analyze the quality of the relations inferred from different types of mappings.

4.3 Small-scale evaluation of the newly inferred relations

The first measure of the quality of the inferred relations is their contribution to retrieval results. However, in order to better understand the retrieval results, we performed a small, in-depth evaluation of the inferred relations and the GTAA-WordNet mappings they were based on. This analysis can point out strengths and weaknesses in our method and guide us in improving the inferred relations and therefore retrieval results.

We evaluated a random sample of 240 inferred relations. Similar to the evaluation of the anchoring, we used a three point scale to quantify our estimation of the

Table 2 Good and bad examples of newly inferred relations and the WordNet relations they were based on

Example relations that we expect to be useful:			
Sidewalk	-	Pavement	*(Meronym 1 step)*
Sand	-	Concrete	*(Meronym 1 step)*
Watch	-	Carillon	*(Hyponym 1 step)*
Electric engine	-	Trolley bus	*(Meronym 1 step)*
Pearls	-	Jewelery	*(Hyponym 3 steps)*
Fjords	-	Seas	*(Meronym 1 step)*
Flour	-	Meal	*(Meronym 3 steps)*
Squid	-	Colouring (pigment)	*(Hyponym 1 step)*
Pharmacy	-	Medicine	*(Hyponym 1 step)*
Barbecues	-	Picknicks	*(Hyponym 2 steps)*
National anthems	-	Music	*(Hyponym 3 steps)*
Pearls	-	Jewelery	*(Hyponym 3 steps)*
Queens	-	Aristocrats	*(Mixed 3 steps)*
Examples of incorrect relations:			
Acupuncture	-	Negotiation	*(Hyponym 1 step)*
Banknotes	-	Copies	*(Hyponym 1 step)*
Apples	-	Foetusses	*(Hyponym 2 step)*

usefulness of the relations: "probably helpful to a wide range of queries" (1), "maybe helpful to some queries" (0.5), "probably not helpful to most queries" (0).

One inferred relation between a pair of GTAA terms is based on two matches to WordNet, one for each GTAA term. A match to WordNet is based on one of more types of lexical information, such as preferred terms or singular forms Table 3 shows the number of inferred relations based on each combination of mapping types. In addition, it shows the number of relations that was in our random sample, and their mean evaluation scores. Note that the sum of all relations in the table is more than the total number of inferred relations (3735), as some mappings were found in more than one way (e.g. based on a singular form and a split compound term). Row or column totals are not meaningfull for this table, as one inferred relation might contribute to more than one row or column. The evaluation scores of inferred relations that were based on two mappings of the same kind are highlighted in boldface; for example, relations based on two preferred terms score 0.50. The number of combinations of two DSSIM mappings in our sample is too small to lead to valid conclusions, and its score is therefore presented in italics.

Table 3 Inferred relations and the combination of mapping relations they were based on: number of inferred relation (T#), number of relations in our evaluation sample (S#) and mean evaluation score (Pr)

Rel.	Pref. terms			Sing. forms			Synonyms			Split compounds			DSSIM		
	T#	S#	Pr	T#	S#	Pr	T#	S#	Pr	T#	S#	Pr	T#	S#	Pr
Pref.	68	10	**0.50**	262	14	0.46	107	8	0.19	598	38	0.34	88	8	0.50
Sing.	262	14	0.46	380	21	**0.52**	378	33	0.23	706	35	0.34	168	23	0.46
Syn.	107	8	0.19	378	33	0.23	96	13	**0.15**	338	21	0.26	81	8	0.19
Split	598	38	0.34	706	35	0.34	338	21	0.26	339	19	**0.18**	533	38	0.25
DSSIM	88	8	0.50	168	23	0.46	81	8	0.19	533	38	0.25	20	2	*0.25*

The relations inferred from mappings based on combinations of preferred terms, singular forms and DSSIM mappings score high (between 0.46 and 0.52), which is in line with the high precision of those mapping types. Relations inferred from mappings based on only synonyms and split compound terms score clearly lower (0.15 and 0.17 respectively), which is also expected given the low scores of those mapping types. However, while synonyms score badly also in combination with preferred terms and singular forms (0.19 and 0.23), the scores of split compounds are less bad when used in combination with preferred terms or singular forms (0.34 for both). These combinations also provide more relations than any other category. This again strengthens our belief that an investigation of an effective and more selective use of split compound terms is a promising direction for both anchoring and thesaurus enrichment.

5 Retrieval with the enriched thesaurus

We employed the enriched thesaurus for retrieval of television programs from the archives of the Netherlands Institute for Sound & Vision. The programs were annotated with subject terms from the GTAA. Our aim is twofold. First, we want to know the value of the inferred relations for retrieval, and compare that to retrieval with existing GTAA relations. Second, we are interested in the added value of the inferred relations when we use them in combination with the existing GTAA relations.

5.1 Experimental setup

We query the collection in nine runs, each using a different type of relation or combination of relations:

Exact	Only programs annotated with the query term are returned. This run is used as a baseline.
GTAA bro	Programs annotated with the query term or broader terms are returned.
GTAA nar	Programs annotated with the query term or narrower terms are returned.
GTAA rel	Programs annotated with the query term or related terms are returned.
GTAA all	Programs annotated with the query term or terms that are related through (a combination of) GTAA relations (narrower, broader, related) are returned.
Via WN 1 step	Programs annotated with the query term or terms related through a one-step inferred relation are returned.
Via WN 2 step	Programs annotated with the query term or terms that are related through a two-step inferred relation are returned.
Via WN 3 step	Programs annotated with the query term or terms that are related through a three-step inferred relation are returned.
Via WN all	Programs annotated with the query term or with a term that is related through a (combination of) one-, two- or three-step inferred relations are returned.

All relations Programs annotated with the query term or terms that are related in any of the above ways are returned.

At present, we allowed three steps between the query term and the target term. More than three steps resulted in an explosion of the number of returned documents.

Of each run, we measure the precision (Prec), recall (Rec) and the harmonic mean of the two, called F_1-measure:

$$\text{Prec} = \frac{|\text{Retrieved\&Relevant}|}{|\text{Retrieved}|} \quad \text{Rec} = \frac{|\text{Retrieved\&Relevant}|}{|\text{Relevant}|}$$

$$F_1 = 2 \cdot \frac{\text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}$$

where |Retrieved| is the number of programs a run retrieved, and |Relevant| is the number of programs that is relevant for a query.

5.2 TRECVID data: corpus and queries

In order to determine the added value of the inferred relations, a dataset and a ground truth are needed that are large enough to distinguish if there are any significant differences between runs. In the current study, we used the TRECVID 2007 development set for the high-level feature extraction task. This dataset consists of 50 hours of news magazine, science news, news reports, documentaries, educational programming, and archival video from the Netherlands Institute for Sound & Vision, 36 queries ('features') and a manually constructed ground truth. A list of the queries can be found in the Appendix at the end of this paper. Sound & Vision kindly provided us with the metadata of this dataset in the form of manual annotations of the television programs with GTAA terms.

The queries consist of a single or moderately complex query term, such as **Sports** or **Explosion_Fire**. This corresponds to the types of queries that are posed in the online search interface of Sound & Vision, where the majority of queries consist of a single term, sometimes completed with a broadcast date. Simple, unequivocal queries are a requirement in this type of study, as complex queries could obscure the results.

We manually translated the high-level features to get queries in terms of GTAA subjects. Features that consisted of two subjects were interpreted as the union of both and we queried for programs containing one and/or the other. This was clearly the intended semantics of the features as can be seen from descriptions such as the one for **Walking_Running**: `Shots depicting a person walking or running`. Of the initial 36 features, three did not have a satisfactory translation, and were therefore discarded.

All TRECVID tasks are at the level of shots, while the GTAA subject annotations are at the level of television programs. We adapted the given ground truth to be on program level; a program was considered relevant to a query if it contained at least one relevant shot. In the resulting ground truth, nine queries appeared in more than 2/3 of the programs and were therefore discarded. `Person` and `Face`, for example, appeared in each program. Six programs were not usable in the present experiment since they did not have a subject annotation and could therefore never be retrieved.

Table 4 Precision, recall and F_1-measure of the nine runs, summarized by the mean ± the standard deviation

Run	Precision	Recall	F_1
GTAA exact	1.00 ± 0.00	0.03 ± 0.06	0.18 ± 0.11
GTAA broader	0.62 ± 0.46	0.03 ± 0.06	0.18 ± 0.10
GTAA narrower	0.94 ± 0.17	0.05 ± 0.08	0.21 ± 0.14
GTAA related	0.38 ± 0.22	0.38 ± 0.21	0.31 ± 0.14
GTAA all	0.33 ± 0.22	0.57 ± 0.24	0.35 ± 0.16
Via WN one-step	0.70 ± 0.32	0.06 ± 0.08	0.21 ± 0.10
Via WN two-step	0.50 ± 0.32	0.07 ± 0.11	0.22 ± 0.10
Via WN three-step	0.54 ± 0.35	0.08 ± 0.12	0.20 ± 0.12
Via WN all	0.37 ± 0.24	0.31 ± 0.29	0.32 ± 0.09
All	0.28 ± 0.19	0.84 ± 0.16	0.37 ± 0.18

After adaptation, the dataset consisted of 104 television programs annotated with on average 3.6 GTAA subject terms, 25 queries and a ground truth that listed on average 27 relevant programs for each query.

5.3 Results and interpretation

Table 4 and Fig. 2 summarize the results. For a detailed overview of all results per query, we refer to the Appendix at the end of this paper. Please note that although the range of the y-axis of the plots in Fig. 2 differ, the height of the bars represents the same value in all three plots.

Throughout this section we use Student's paired t-test to compare the performance of runs.[11] Significance levels, t-values, degrees of freedom and the appropriate version of the test (one or two tailed) will be reported as, for example, ($t =$, $p =$, $df =$, one-tailed).

5.3.1 Existing GTAA relations

The results of the runs using existing thesaurus relations merely confirm previous findings on thesaurus-based retrieval (e.g. [9, 26]). We discuss them since they form a baseline against which we can compare the performance of the inferred relations. The human entered subject terms are reliable, and using them gives high precision, in our case even 100% (the 'exact' run). We suspect that the level of correctness of our annotations was higher than usual thanks to the special attention the Netherlands Institute for Sound & Vision gave to the collection they prepared for TRECVID. In many cases, of course, human annotators do err and disagree [28]. The time-consuming nature of human annotation causes the number of subject terms per program to be low, much lower than the number of topics that is visible in the

[11] The t-test requires a normal distribution. Normality was assessed with Quantile-Quantile plots. Although for some of the smaller samples—the exact run, for example, returned programs for only seven queries and had therefore only 7 precision values—normality could not be proven, we assume that precision and recall are normally distributed quantities given a large number of queries.

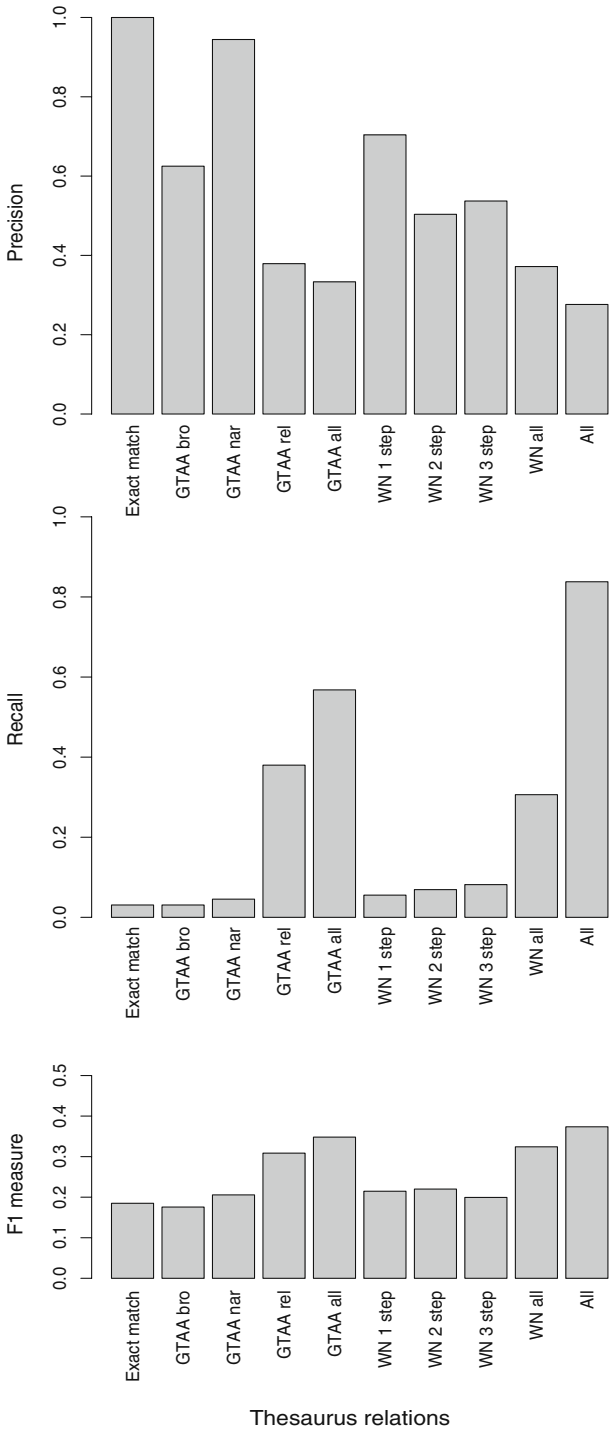

Fig. 2 Retrieval results with different thesaurus relations

video. This makes the recall of the run that relies solely on these human annotations unacceptably low: 3% on average.

Including terms that are broader than the query does not add to recall. This is partly due to the fact that our queries are all fairly general, and many don't have a broader term. Still, it is a confirmation of what was found in an earlier study [9]. Narrower terms, on the other hand, do seem to add a little to recall, although the result is not statistically significant ($t = 1.56$, $p = 0.07$, $df = 23$, one-sided) and they maintain a high precision. This is what we would expect from the definition of narrower terms: "the scope (meaning) of one falls completely within the scope of the other" [19]. Related terms are less reliable: precision drops to just over one-third compared to using only exact matches ($t = 7.33$, $p < 0.01$, $df = 6$, two-sided), but recall increases to 38% ($t = 8.21$, $p < 0.01$, $df = 23$, one-sided).

Combining the hierarchical broader/narrower relations with the related terms (the "GTAA all" run), only slightly (but significantly) lowers precision further compared to using only the related terms ($t = 2.6$, $p = 0.02$, $df = 23$, two-sided). It does, however, raise recall to 57% ($t = 6.71$, $p < 0.01$, $df = 23$, one-sided). This suggests that also sequences of different types of relations are beneficial to retrieval.

5.3.2 Newly inferred relations

The one-, two- and three-step inferred relations perform equally well. One-step scores a higher precision, but not significantly so ($t = 2.18$, $p = 0.06$, df = 8, two-sided, compared to two-step; $t = 0.26$, $p = 0.80$, df = 7, two-sided, compared to three-step). Also regarding recall, there were no significant differences between one-, two- and three-step inferred relations. This suggests that the notion of relatedness can be interpreted in a broad sense and does not need to be restricted to only one step in the WordNet hierarchy.

When the one-step, two-step and three step inferred relations are combined (the "Via WordNet all" run), precision falls to 37%. Recall, on the other hand, rises to 0.31%. These results are comparable to the results of existing relations in the GTAA that were created by experts. When comparing the inferred relations (the 'Via WordNet all' run) to GTAA related terms we observe that they have similar precision scores, but somewhat lower recall. The difference in recall can in part be explained from the fact that there are twice as many related terms as inferred relations. With respect to F_1-measures, there was no significant difference between the inferred relations and GTAA related terms ($t = 0.7$, $p = 0.50$, df = 15, two-sided). When comparing the inferred relations to GTAA narrower terms, they score better on recall but worse on precision, resulting in a significantly higher F_1-measure ($t = 2.7$, $p = 0.03$, df = 8, two-sided). These results suggest that the inferred relations are valuable for retrieval in situations where there is no other structure in the vocabulary. In these cases, they could serve the same purpose as related terms and, to a lesser extend, narrower terms.

Using all relations together improves the recall significantly over using only the existing GTAA relations ($t = 7.8$, $p < 0.01$, $df = 23$, one-sided). This suggests that enrichment of a weakly structured thesaurus has added value to the retrieval results. In addition, it suggests that the combination of different types of relations is

beneficial to the retrieval results: recall increases and precision decreases, resulting in an alltogether positive effect on the F_1-measure. This phenomenon was also observed when comparing the use of all GTAA relations to only one type of GTAA relation, and comparing the use of all WordNet relations to only one-, two- or three-step WordNet relations. In these situations, the increase in recall could in part be attributed to the higher number of retrieved programs. We calculated the increase in recall that we would expect if the additionally retrieved programs were randomly taken from the collection, as follows:

$$\mathbb{E}_{incr} = \frac{(R_{combi} - R_{one}) \cdot (C - RC_{one})}{N - R_{one}} \cdot \frac{1}{C}$$

where $\mathbb{E}_{incr}$ is the expected increase in recall for a query, R_{combi} and R_{one} are the number of retrieved programs in the two compared runs, the first using a combination of relations and the second using only one type of relation. C is the number of correct programs for the query in the collection, RC_{one} is the number of correctly retrieved programs in the run using only one type of relation and N is the total number of programs in the collection (104 in our case).

We compared the expected to the observed increase in recall using Student's t-test for (1) the "GTAA all" run and the runs using one type of GTAA relation, (2) the "Via WordNet all" run and runs using one type of WordNet relation and (3) the "All relations" run and the "GTAA all" run. In all cases the differences between expected and observed increase were significant at the 0.01 α-level. However, the latter needs a closer inspection. The mean increase in recall from the "GTAA all" run to the "All relations" run was 0.27. The mean expected increase in recall was 0.21, which is significantly lower than the observed increase in recall ($t = 3.18$, $p < 0.01$, $df = 23$, two-sided). Still, this means that a large portion (roughly 3/4) of the observed increase in recall (0.27) can be attributed to the higher number of retrieved programs. As can be seen from the Appendix, the number of retrieved programs are exceptionally large in the "All relations" run. This shows that the choice to allow three steps between query and annotation (see Section 5.1) does work out in combination with a retrieval strategy that uses all possible combinations of relations.

5.4 Towards an interdisciplinary comparison of results

Although we have used a TRECVID dataset in the experiments, our approach is very different from the approach of systems that participate in the TRECVID conference. Firstly, our approach is based on metadata and the structure of a thesaurus, while TRECVID retrieval systems are based on the audiovisual signal. The latter can include speech-to-text transcriptions but, as was shown in [18], these result in textual descriptions that are different from keywords manually assigned by cataloguers. Secondly, we retrieve complete television programs, while TRECVID participants retrieve individual shots in a program. Also with respect to the evaluation method the differences are considerable: TRECVID uses (Inferred) Average Precision (AP) to evaluate a ranked result list where we use Precision, Recall and F_1-measure to evaluate a unordered set of results. Still, a comparison between two disciplines that work on the same dataset is a valuable exercise that puts the results in a broader perspective. We performed a qualitative examination in which we place the scores of the TRECVID 2007 feature task next to our results. Because a direct comparison of

Table 5 Ranks of TRECVID average AP scores, and ranks of the F_1-measure of metadata-based retrieval (the "All relations" run.) on 18 queries

Query	Mean TRECVID rank	"All relations" rank
Waterscape_Waterfront	1	3
Meeting	2	7
Car	3	1
Animal	4	4
Maps	5	13
Airplane	6	17
Charts	7	14
Office	8	2
Boat_Ship	9	9
Desert	10	18
Sports	11	12
Mountain	12	8
Truck	13	5
People-marching	14	11
Military	15	10
Explosion_Fire	16	15
Weather	17	6
Police_Security	18	16

the AP scores to our scores is not meaningful, we compare the *ranks* of each query (feature) instead.

We have evaluated our approach on the TRECVID development set, which has a ground truth for 36 queries (features). Three queries did not have a satisfactory translation in the GTAA, and were therefore discarded. TRECVID participants, on the other hand, used the development set to develop their systems, and were evaluated on the test set which has a ground truth for only 20 of the 36 queries. We can only compare the queries on which both approaches were evaluated. These queries are enumerated in Table 5. They are ranked according to the mean Inferred AP score of all TRECVID 2007 participants, where 1 represents the highest score. The ranks of the F_1-measures of the "All relations" run in our approach are given in the third column. When comparing the ranks, we assume that they reflect the difficulties that the query poses to each approach. However, it should be noted that the rank of the TRECVID systems is also determined by the prior chance to find a concept in the collection: if a concept appears few shots, it is hard to reach a high AP, while if a concept appears in many shots, a high AP score is within easy reach.

When inspecting the ranks, there seems to be a weak correlation between the ranks of the two approaches, but we could not prove this statistically (Spearman's $r = 0.37$, $p = 0.13$). Interestingly, there is also a certain complementarity of the approaches: on concepts with a clear visual appearance like *Maps* and *Charts*, the TRECVID systems get high scores, while on concepts that are more abstract, like *Weather*, our metadata-based approach scores high. This follows the intuition that, indeed, Maps and Charts will hardly ever be mentioned in the description of a TV program, unless it is actually about maps and charts. The weather and related concepts such as rain, wind, climate and cold, are topics much discussed in the Netherlands, which could explain the good coverage of the metadata. A deciding factor in the quality of a machine learned concept detector as used in TRECVID is the number and range of the training examples. The concept Weather has an

unlimited amount of visual appearances, and it seems unlikely that a good coverage of these appearances can be realized in a training set to develop reliable weather detectors, which would explain the poor performance. The complementarity of semantics-based and signal-based methods was also noted in [24].

Nevertheless, it is hard to define the characteristics of a concept that make it more suitable for one of the two approaches. For example, the concept *Car* appears in the top ranks of both methods. *Airplane* is retrieved well by the average TRECVID system, but not by our metadata based approach. For *Truck* it is the other way around. More research is necessary on how the two approaches can be used as alternative or complementary retrieval methods.

6 Discussion and future work

In this paper we experimented with retrieval using a thesaurus that was enriched by anchoring it to an external resource. We have shown that with simple techniques new relations can be inferred that are valuable for retrieval purposes. This creates possibilities to improve retrieval in the Sound & Vision collection and other collections indexed with unstructured or weakly structured thesauri.

We investigated both the effect of only using the newly inferred relations and using them in combination with existing thesaurus relations. Retrieval with only the inferred relations yielded an F_1-measure of around 0.3. This is comparable to the intuitive approach of using `Related Term` thesaurus relations. This finding suggests that it is possible to use an external resource to enrich an otherwise unstructured vocabulary and base the retrieval on the inferred relations. For example, we see possibilities to enrich lexicons of high-level feature detectors that are used in content-based video retrieval. LSCOM is such a vocabulary for annotation and retrieval of video, containing concepts that represent realistic video retrieval problems, are observable and are (or will be) detectable with content-based video retrieval techniques [22]. In a recent effort, LSCOM was manually linked to the CyC knowledge base,[12] thus creating structure within LSCOM. We would be interested in comparing and combining this manually added structure with the enrichment that would be the outcome of the methods proposed in the present paper.

When the inferred relations are used in combination with existing thesaurus relations, it appears that the use of the enriched set of relations increases recall (from 0.57 in the "GTAA all" run to to 0.84 in the "All relations" run) with only a slight drop in precision (0.33 vs. 0.28). This indicates that it is beneficial to enrich an already structured thesaurus. However, a larger test collection will be needed to confirm this finding.

When looking at the F-measure scores it is interesting to note that mixing relationships increases performance, both in the non-enriched ("GTAA all") and in the enriched ("All relations") case. This suggests that the nature of the relationship (broader, narrower, related) is not a big issue, at least not in this case. It would be worthwhile to study this in more detail in future experiments.

[12] http://www.cyc.com/

The next step in this line of research would be to investigate the use of newly inferred relations in the ranking of results. We could, for example, set a 'traversal cost' for each type of relation, as in [26]. A related issue is the adaptation of the groundtruth from shot level to program level. A ranked list of relevant programs, based on the number of relevant shots they contain, could be used to evaluate a ranked list of results.

Our thesaurus enrichment approach was based on an anchoring to WordNet. In a small evaluation study, we have shown that different types of anchoring lead to differences in the quality of the inferred relations. We were particularly intrigued by the relations that were inferred from a mapping based on the split parts of a compound term. This type of mapping seems to be especially promising when a mapping based on the complete term is not found. However, the high number of incorrect relations that were inferred from this type of mapping, calls for more research on how to successfully put to use the split compound terms for thesaurus enrichment. In addition, further experiments could reveal what the effect of different WordNet relations is on our thesaurus enrichment approach: hyponyms, meronyms, but also other relations that were not yet used. Finally, we are interested to compare the outcome of thesaurus enrichment methods that do not make use of explicit semantic knowledge, such as approaches using Hearst-like patterns to extract relations between terms from text or from the web.

The use of TRECVID data enabled us to experiment on a dataset of reasonable size. However, it also raises some issues. The TRECVID ground truth is based on the pooled results of TRECVID 2007 participants: only items returned by at least one of the participants are judged, while items not returned by anyone are considered incorrect. This could in theory lead to a negative image of our results, since we did not contribute to the pool. It is been argued that this is a negligible problem. Zobel [30], for example, demonstrated that the difference in rating between a system in- and outside the pool is small. However, all systems in his test were content-based image retrieval systems. The retrieval approach under consideration in the present paper is concept-based. Since we use another type of information (metadata instead of the audiovisual signal), it is well possible that we retrieve a set of documents that is disjoint from the set that was retrieved by the content-based systems that contributed to the pool. Therefore, the effect of being outside the pool is potentially larger.

The translation of TRECVID topics to GTAA terms was done manually. In a final application this translation would be done either automatically, which is done in [24], or by the searcher, as in [10]. However, in the present paper our goal was not to build an application but to investigate the possibilities of retrieval with an automatically enriched thesaurus.

Acknowledgements This work was carried out in the context of the MuNCH and CHOICE projects, which are supported by the Netherlands Organisation for Scientific Research (NWO) programme for Continuous Access To Cultural Heritage (CATCH) under project numbers 640.002.501 and 640.001.402.

Appendix

Table 6 Number of relevant documents (Rel. docs), retrieved document (Ret) and number of relevantly retrieved documents (Rel) for each query by each run

Query	Rel. docs	Exact		GTAA								WordNet								All	
				Narrower		Broader		Related		All		One step		Two steps		3 steps		All			
		Ret	Rel	Ret	Rel	Ret	Rel	Ret	Rel	Ret	Rel	Ret	Rel	Ret	Rel	Ret	Rel	Ret	Rel	Ret	Rel
Sports	26	1	1	1	1	4	4	65	16	81	23	2	2	6	2	1	1	30	12	93	23
Office	58	0	0	0	0	0	0	7	5	57	36	0	0	0	0	0	0	0	0	66	42
Meeting	38	0	0	0	0	0	0	12	9	14	10	19	3	0	0	0	0	48	13	76	26
Studio	11	0	0	0	0	0	0	23	7	46	7	0	0	0	0	0	0	0	0	48	7
Desert	5	1	1	2	1	1	1	9	2	33	3	1	1	1	1	6	2	28	4	64	5
Mountain	23	0	0	1	0	0	0	5	1	18	8	0	0	0	0	0	0	0	0	46	15
Snow	8	0	0	0	0	0	0	6	2	14	3	0	0	5	1	1	0	25	4	60	8
Military	24	3	3	3	3	3	3	40	13	66	18	3	3	3	3	3	3	3	3	84	23
Prisoner	6	0	0	0	0	0	0	30	4	47	5	0	0	0	0	0	0	0	0	83	6
Animal	45	3	3	3	3	12	12	43	22	67	29	19	12	11	6	19	13	65	32	99	43
Airplane	11	0	0	0	0	0	0	19	4	39	5	4	1	19	4	7	2	69	10	94	11
Car	68	0	0	0	0	2	1	37	23	67	44	0	0	0	0	11	5	34	19	89	58
Bus	19	1	1	1	1	1	1	14	5	22	8	4	2	2	1	1	1	37	10	77	15
Truck	52	0	0	0	0	0	0	10	7	18	12	0	0	0	0	11	4	36	15	73	34
Boat_Ship	29	0	0	0	0	1	1	23	11	65	19	0	0	1	0	5	1	24	8	97	27
Natural-disaster	5	0	0	0	0	0	0	16	4	32	4	0	0	0	0	0	0	0	0	56	5
Maps	27	0	0	0	0	0	0	6	2	15	5	0	0	0	0	0	0	0	0	46	13
Charts	26	0	0	1	0	0	0	12	2	18	4	0	0	0	0	9	2	30	9	71	16
Waterscape_Waterfront	65	3	3	3	3	3	3	24	20	39	30	6	5	4	3	6	4	14	11	61	42
Weather	31	0	0	1	0	0	0	35	15	60	20	0	0	0	0	0	0	0	0	75	25
Court	2	0	0	0	0	0	0	21	1	61	2	0	0	0	0	0	0	0	0	69	2
Police_Security	15	3	3	4	3	3	3	39	8	82	13	3	3	7	4	3	3	26	6	96	15
People-marching	23	0	0	0	0	0	0	27	14	40	17	3	2	14	6	0	0	56	14	87	22
Explosion_Fire	17	0	0	0	0	0	0	24	5	42	7	0	0	0	0	19	6	51	10	89	15

References

1. Assem van M, Malaisé V, Miles A, Schreiber ATh (2006) A method to convert thesauri to skos. In: Proceedings of the third European semantic web conference, pp 95–109. Budvar, Montenegro
2. Broekstra J, Kampman A (2003) SeRQL: a second generation RDF query language. In: Proceedings of the SWAD-Europe workshop on semantic web storage and retrieval, Amsterdam, pp 13–14
3. Caracciolo C, Euzenat J, Hollink L, Ichise R, Isaac A, Malaisé V, Meilicke C, Pane J, Shvaiko P, Stuckenschmidt H, Šváb Zamazal O, Svátek V (2008) Results of the ontology alignment evaluation initiative 2008. In: The third international workshop on ontology matching at ISWC
4. Euzenat J, Isaac A, Meilicke C, Shvaiko P, Stuckenschmidt H, Šváb O, Svátek V, van Hage WR, Yatskevich M (2007) First results of the ontology alignment evaluation initiative 2007. In: Ashpole B, Ehrig M, Euzenat J, Stuckenschmidt H (eds) Ontology matching, CEUR workshop proc
5. Fellbaum C (ed) (1998) WordNet: an electronic lexical database. MIT, Cambridge
6. Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. In: In Proceedings of the 14th international conference on computational linguistics, pp 539–545
7. Hirst G, St-Onge D (1998) Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum C (ed) WordNet: an electronic lexical database. MIT, Cambridge, pp 305–332
8. Hollink L, Malaisé V, Schreiber G (2008) Enriching a thesaurus to improve retrieval of audiovisual documents. In: The third international conference on semantic and digital media technologies (SAMT)
9. Hollink L, Schreiber G, Wielinga B (2007) Patterns of semantic relations to improve image content search. Journal of Web Semantics 5:195–203
10. Hollink L, Schreiber ATh, Wielemaker J, Wielinga BJ (2003) Semantic annotation of image collections. In: Proceedings of the K-Cap 2003 workshop on knowledge markup and semantic annotation
11. International Organization for Standardization (1986) ISO 2788:1986. Guidelines for the establishment and development of monolingual thesauri. ISO, Geneva
12. Kekäläinen J, Järvelin K (2002) Using graded relevance assessments in IR evaluation. J Am Soc Inf Sci Technol 53(13):1120–1129
13. Khan LR, Hovy E (1997) Improving the precision of lexicon-to-ontology alignment algorithm. In: AMTA/SIG-IL first workshop on interlinguas. San Diego, CA, USA
14. Knight K, Luk S (1994) Building a large-scale knowledge base for machine translation. In: The AAAI-94 conference
15. Lesk M (1986) Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: SIGDOC '86: proceedings of the 5th annual international conference on systems documentation. ACM, New York
16. Malaisé V, Zweigenbaum P, Bachimont B (2007) Mining defining contexts to help structuring differential ontologies. In: Application-driven terminology engineering
17. Malaisé V, Isaac A, Gazendam L, Brugmann H (2007) Anchoring dutch cultural heritage thesauri to wordnet: two case studies. In: ACL 2007 workshop on language technology for cultural heritage data. Prague, Czech Republic
18. Malaisé V, Isaac A, Gazendam L, Heeren W, Ordelman R, Brugmann H (2009) Relevance of ASR for the automatic generation of keywords suggestions for TV programs. In: Conférence sur le traitement automatique des langues TALN
19. Miles A, Brickley D (2005) SKOS core guide. W3C working draft electronic document. http://www.w3.org/TR/swbp-skos-core-guide/. Accessed February 2008
20. Miles A, Bechhofer S (2008) SKOS simple knowledge organization system reference. W3C working draft electronic document. http://www.w3.org/TR/skos-reference/. Accessed April 2008
21. Nagy M, Vargas-Vera M, Stolarski P (2008) Dssim results for oaei. In: The third international workshop on ontology matching at ISWC
22. Naphade M, Smith JR, Tesic J, Chang SF, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE MultiMedia 13(3):86–91. http://doi.ieeecomputersociety.org/10.1109/MMUL.2006.63
23. Over P, Kraaij W, Smeaton AF (2007) TRECVID 2007—an introduction. In: TREC Video retrieval evaluation online proceedings
24. Snoek CGM, Huurnink B, Hollink L, de Rijke M, Schreiber G, Worring M (2007) Adding semantics to detectors for video retrieval. IEEE Trans Multimedia 9(5):975–986

25. Snow R, Jurafsky D, Ng AY (2005) Learning syntactic patterns for automatic hypernym discovery. In: Saul LK, Weiss Y, Bottou L (eds) Advances in neural information processing systems, vol 17. MIT, Cambridge, pp 1297–1304
26. Tudhope D, Binding C, Blocks D, Cunliffe D (2006) Query expansion via conceptual distance in thesaurus indexed collections. J Doc 62(4):509–533
27. Van Hage WR, Kolb H, Schreiber G (2006) A method for learning part-whole relations. In: Proceedings of the international semantic web conference (ISWC 2006), pp 723–735
28. Volkmer T, Thom JA, Tahaghoghi SMM (2007) Exploring human judgement of digital imagery. In: ACSC '07: proceedings of the thirtieth Australasian conference on computer science. Australian Computer Society, Darlinghurst, pp 151–160
29. Voorhees E (1994) Query expansion using lexical-semantic relations. In: Croft WB, Rijsbergen van CJ (eds) Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval. Springer, New York, pp 61–69
30. Zobel J (1998) How reliable are the results of large-scale information retrieval experiments? In: SIGIR '98: proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 307–314. http://doi.acm.org/10.1145/290941.291014

Laura Hollink received the M.A. degree in social science informatics from the University of Amsterdam, The Netherlands, in 2001. In 2006 she received the Ph.D. degree in computer science from the VU University, Amsterdam, on the topic of semantic annotation for retrieval of visual resources. Since then, she works in the NWO MuNCH project in close cooperation with the Netherlands Institute for Sound and Vision on retrieval in audiovisual archives using semantic techniques. She has participated in the Multimedian Eculture project on search through distributed collections of cultural heritage objects. Currently, she is an assistant professor at the Web and Media Group, VU University.

Véronique Malaisé is a Postdoctoral researcher at the VU University of Amsterdam, in the Intelligent Information Systems Group. She graduated in 2005 in Linguistics, with a specialisation in Natural Language Processing, from the University of Paris VII. She then obtained a Postdoc position in the NWO CHOICE project, working in collaboration with the Netherlands Institute for Sound and Vision, the Dutch National Audiovisual Archives. This research topic was the follow up of her PhD research, centered around ontologies and Audiovisual documents annotation. She is currently working on a Dutch project aiming at combining low-level features with semantic representation, but this time in the maritime domain: the Poseidon project.

Guus Schreiber is a professor of Intelligent Information Systems at the Department of Computer Science department of the VU University Amsterdam. His research interests are mainly in knowledge and ontology engineering, with a special interest for applications in the field of cultural heritage. He was one of the key developers of the CommonKADS methodology. He acts as chair of W3C groups for Semantic Web standards such as OWL, SKOS and RDFa. His research group is involved a wide range of national and international research projects. He is now project coordinator of the EU Integrated Project NoTube concerned with integration of Web and TV data with the help of semantics and was previously Scientific Director of the EU Network of Excellence "Knowledge Web".