Improving the Accessibility of a Thesaurus-Based Catalog by Web Content Mining

Gijs Geleijnse Jan Korst

Philips Research
High Tech Campus 34, 5656 AE Eindhoven, The Netherlands
{gijs.geleijnse,jan.korst}@philips.com

Abstract. In this work we focus on the improvement of the accessibility of a catalog of radio and television productions. As productions can only be searched using the annotated meta-data, the use of a controlled vocabulary plays an important role in the retrieval. However, users can not be expected to have detailed knowledge on the terms defined within the vocabulary. In this work, we present a method that assists users to find appropriate keywords for their queries, making the archive better accessible both for professional users and for the general public. The experimental results show that the algorithm developed can be of assistance for those working with the thesaurus.

1 Introduction

To consistently annotate items and to facilitate their retrieval, cultural heritage institutions often use controlled vocabularies for indexing their collections. For the Netherlands Institute for Sound & Vision (S & V)[1], the annotations are currently the only basis to retrieve the audiovisual content. S & V uses a dedicated thesaurus, the GTAA[2], as a controlled vocabulary. Especially for the items that are briefly annotated, e.g. where a summary of the content is missing, searching for GTAA terms is the most effective mechanism for retrieval.

Although the use of a controlled vocabulary such as the GTAA provides a uniform annotation over the whole collection, it also gives rise to two problems. On the one hand, the retrieval of items – both for professionals and for the general public – depends on the knowledge of the content of the GTAA. On the other hand, the controlled vocabulary needs to be updated from time to time as new terms become relevant.

In this paper we address the following problem. Given an arbitrary term, we are interested in the term within the thesaurus that is semantically closest. Such a mapping between the term and the GTAA can be of assistance for those who want to search the catalog as it will provide more effective search results. For those annotating an audiovisual production it can also be of use, as it can help to find the closest terms within the GTAA.

[1] http://www.beeldengeluid.nl

[2] Gemeenschappelijke Thesaurus Audiovisuele Archieven (*Common Thesaurus Audiovisual Archives*).

For many languages, such as Dutch, no structured knowledge is available to derive a mapping between an arbitrary term and the thesaurus. We therefore use unstructured texts to extract such a mapping, by deploying techniques developed in the fields of ontology mapping and web content mining. We derive semantic relations between a query term and the thesaurus using search engine snippets.

We illustrate that the method presented is domain and language independent by evaluating both a mapping of terms to the Dutch GTAA and the Agricultural Thesaurus[3] (NALT) of the US National Agricultural Library.

2 Related Work

Together with the development of the Semantic Web, the research topic of Ontology Matching arose [1]. In Ontology Matching, the task is to combine or create relations between two separately designed ontologies. Although most approaches are based on the structures of the ontologies combined with lexical matches (e.g. [2–4]), the use of web content mining has recently been deployed for this task ([5, 6]).

The use of text patterns to extract hyponym relations from unstructured texts was coined by Hearst [7, 8]. She observed that patterns like *such as* and *for example* linked terms with their hypernyms (i.e. their broader terms). Previous work shows that such patterns can be used to extract thesaurus-relations from the web using a search engine e.g. [9–13]. Given a training set of hyponym-hypernym pairs, these patterns can also be learned automatically [14, 15].

Web information extraction applied to the cultural heritage domain is addressed by De Boer et al. [16]. Here, RDF ontologies of painters and art movements are linked by analyzing web pages on art movements. The numbers of search engine hits is used in [17] to identify the periods corresponding to art styles. In [18] such numbers are used to identify relatedness between Dutch 17th century painters, where in [19] a 'social network' is computed for historical figures.

As an alternative approach to the use of web content mining to improve the accessibility of the catalog, Malaisé et al. [20] created a method to link the Dutch GTAA thesaurus to the English WordNet [21] via a bilingual online dictionary. As the GTAA contains many multi-word terms and compounds, such a mapping can not always be found. Moreover, it is not trivial to link an arbitrary given term via WordNet to the GTAA.

3 Problem Description and Outline

Given is a thesaurus, i.e. a list of terms and their semantic relations. Typical relations are the *broader term* relation (BT) between a term and a more general term (e.g. *herring gull* and *seagull*), its counterpart the *narrower term* relation (NT), and the *related term* RT relation for terms that are associated with one another. Moreover, a thesaurus can contain preferred and non-preferred term. The latter refer to the first via the *use* relation (US), *used for* (UF) is its inverse. For each preferred term, the GTAA also distinguishes

[3] `http://agclass.nal.usda.gov/agt/agt.shtml`

bioscooppersoneel *(cinema personnel)*	earthworms
1D05.03 *economy – trades, services*	BT: invertebrates
1D12.01 *arts and culture – general*	BT: soil invertebrates
1D13.02 *sports and leisure – recreation*	RT: earthworm burrows
BT personeel *(personnel)*	RT: Lumbricidae
BT werknemers *(employees)*	RT: vermiculture
NT filmoperateurs *(film operators)*	RT: worm casts
RT bioscopen *(cinemas)*	
RT film *(film)*	
UF explicateurs (± *silent film commentator)*	

Table 1. Example terms from the GTAA (left) and the NALT (right)

one or more categories, subdivided into 15 main categories (e.g. *sports and leisure*) and each with 3 to 7 subcategories (e.g. *recreation*). The terms in the GTAA are mostly in plural, but the singular forms are added as well.

Example terms from both the GTAA and the NALT are given in Table 1. For the GTAA term the translations in English are given. For example, the entry shows that *invertebrates* is a broader term (BT) for *earthworms*.

Currently, detailed knowledge of the content of the GTAA thesaurus T is crucial for describing (and redescribing) items within the catalog. Moreover, the recall of briefly described items will improve when using search terms within the GTAA. Hence, an assistant is desired that suggests terms from the GTAA for a given query term. Ideally, for a given term v we are interested in a synonym of v within T. As the GTAA only consists of about 5,000 terms, it is not likely that a synonym is present for v. We therefore focus on finding the narrowest broader term t for v. For example, if we are interested in a mapping for the term *albatross*, the terms *bird* and *animal* are indeed broader terms, but are too broad. *Seabird* however would be the narrowest broader term for *albatross*.

The algorithm presented is to be used as an assistant. For a given query v, we therefore present multiple candidate terms with hyperlinks to the GTAA. Even if t is not identified by the assistant, t can easily be found if the method does return terms that are semantically close (e.g. by the RT relation) and hence links to t. Instead of navigating through a thesaurus with 5,000 terms, the user is now only presented a handful of alternatives. Hence, the user can select a term at a single glace.

3.1 Global Outline

As no suitable structured information is available for this task, we use unstructured texts found with a search engine to determine a mapping from v to the thesaurus. Using the *Yahoo!* API for our experiments, we are allowed to perform 5,000 automated queries a day. Approaches as discussed in e.g. [9, 18] have a query complexity of $\mathcal{O}(|T|)$ per term. We therefore aim at an approach more efficient in the number of queries per term.

As a first approximation, we start with determining the most relevant categories (Table 1) for v (Section 4). We use the computed categories in the next steps: three alternative approaches in mapping v to T. In Section 5 we discuss the pattern-based

method used to find the mapping. In Section 6 we derive an alternative mapping using enumeration patterns. As lexically similar terms rarely occur in the same sentence, we use a simple lexical method to find the mapping in Section 7. Section 8 handles the combination and presentation of the results of the three mapping techniques. The experiments are presented in Section 9, while we conclude in Section 10.

4 Determining a Category

A commonly used paradigm in natural language processing is that the semantics of a term can be determined by its discourse [22]. We use this assumption to first determine the category of the term v to be mapped to the thesaurus T. For each category r, we compute a score $s_v(r)$.

We collect the 100 snippets for the query "v". We choose to use only the snippets, as downloading the full documents is too time consuming. Earlier work shows that good results can be obtained when using snippets for web information extraction [6, 23].

Having collected the snippets, we scan them for terms in T. Each term in T that occurs in the snippets contributes to the scores of its categories. Hence, if the term *bioscooppersoneel* (see Table 1) occurs in the snippets found with v, this occurrence contributes to the scores of the categories 1D05.03, 1D12.01 and 1D13.02.

As infrequent terms are more discriminative than frequent ones (e.g. *people*), the occurrences of the terms in T are weighted by their estimated total frequency on the web. Words such as *haar* (either *hair* or *her* in English) that appear frequently in Dutch texts get a lower score than infrequent terms such as *1 mei-vieringen* (May 1 celebrations).

The score for category r is computed as follows.

$$s_v(r) = \sum_{t \in r} \text{oc}(t) \cdot \log \frac{C}{f(t)} \tag{1}$$

Where C, $f(t)$, and $\text{oc}(t)$ are defined as such:
$\text{oc}(t) =$ the number of times term t (or its singular form) occurs ,
$f(t) =$ the number of search engine hits for the query "t", and
$C = \sum_{t \in T} f(t)$ i.e. sum of all hits.

After having computed the scores for each of the categories, we assign the category r_j with the highest score to v. As a term can be within multiple categories, we also add the categories with at least half the score of r_j. Hence, we add all categories for which the following holds.

$$s_v(r_i) \geq 0.5 \cdot \max_j s_v(r_j) \tag{2}$$

We will use the categories in the thesaurus mapping techniques described in the next three sections.

5 Term Mapping using hyponym patterns

In earlier work [15] we describe how effective relation patterns can be learned using a training set of pairs of related terms. As this is beyond the scope of this paper, we assume a set of patterns to be given that relate Dutch terms with their hypernyms [7]. IJzereef manually constructed such a set in [24].

Having a set of patterns, we combine the term v with each of the patterns into queries. For example, suppose that we are interested in the term *puffins* and *such as* is one of our patterns. We then query the expression "such as puffins" and search for terms in T preceding the search term. Hence, the aim is to find phrases like *'seabirds such as puffins'* to determine broader terms for *puffins*. For efficiency reasons, we only scan the snippets returned by the search engine.

For a term $t \in T$ found within the snippets for query term v, we compute its score $s_v(t)$ as follows.

$$s_v(t) = q(t, v) \cdot \text{oc}(t) \cdot \log \frac{C}{f(t)} \tag{3}$$

We use $q(t, v)$ as a penalty score for terms outside the subcategories found in the previous section.

$$q(t, v) = \begin{cases} 1.0 \text{ if } t \text{ and } v \text{ share a subcategory} \\ 0.3 \text{ if } t \text{ and } v \text{ share a main category} \\ 0.1 \text{ if } t \text{ and } v \text{ share no category} \end{cases}$$

The values for $q(t, v)$ are now chosen in a somewhat arbitrary way. In future work, optimal values for these scores are to be determined empirically.

Using the scores, we compute a ranked list for the potential hypernym terms for v found using this method.

6 Term Mapping using Enumeration Patterns

Snow et al. observe that related terms (or *siblings*) tend to co-occur in enumerations [25, 26]. We thus can state that enumerated items share a broader term. Hence, if we can observe which terms within T are siblings of v, we can use the structure of the thesaurus to compute the broader term for v.

Similar to the approach described in the previous section, we select a number of patterns expressing the RT relation. Again, we scan the snippets for terms within the thesaurus. However, we do not score the terms found, but (all) their broader terms. Hence, the presence of the term *aalscholvers* (cormorants) contributes to the scores for *watervogels* (water birds), *vogels* (birds), and *dieren* (animals).

A term t is hence scored using the presence of all its narrower terms $\text{NT}^*(t)$ in the snippets.

$$s_v(t) = \sum_{s \in \text{NT}^*(t)} q(s, v) \cdot \text{oc}(s) \cdot \log \frac{C}{f(s)} \tag{4}$$

We assume that the broadest concepts (e.g. *dieren*, animals – 26,000,000 hits) are in general more present on the web than narrower concepts (e.g. *watervogels*, waterbirds – 230,000 hits). Hence, we do take the distance of s to t into account as the factor $\frac{C}{f(s)}$ penalizes common concepts. Again, we compute a ranked list of potential hypernym terms using this enumeration-based approach.

7 Term Mapping using a Lexical Approach

We observe that hyponym-hypernym pairs that are lexically similar (e.g. *dienstverlenende beroepen* and *beroepen*, *earthworms* and *worms*) occur infrequently within the same sentence. Next to the two approaches based on web information extraction, we therefore adopt an approach using the morphology of the terms.

If some term t in T is a suffix of v, then v may be a hypernym of t (e.g. if v contains a preceding adjective). However, not all t that match with a suffix of v are indeed hypernyms of v. For example, the GTAA term *ogen* (eyes) is a suffix of psychologen (psychologists).

However, if the computed categories for v do not overlap with the categories for suffix t, it is not likely that the two are related. We therefore use the categories as computed in Section 4 to filter out erroneous lexical mappings.

We construct a list of thesaurus terms that are suffixes of v and share a sub category with v. If no such terms exist, we create such a list of terms that share a main category with v. The list is sorted by increasing length.

8 Presenting the results

Having independently found three lists of potentially relevant terms for the query v, the task is to identify a mapping, i.e. the most relevant term in T. We search the three lists to select the '*best of three*'.

This *best of three* term is selected as follows. We select the term with the highest average rank that is found by all methods. If no term exists, we select the term with the highest average rank over two of the three methods. We leave this '*best of three*' field blank if no term is identified by more than one method.

As the algorithm presented is intended to be an assistant for the user of the catalog rather than a fully automatic mapping, we also present the outputs of the individual methods. An HTML page is generated where the terms are linked with the corresponding entries in the thesaurus. Hence, even if the best suited term is not found, the user can navigate to this term by clicking a closely related term.

As the number of queries is linear in the number of patterns for a given input term, a real-time application is possible. With an (inefficient) implementation where 21 queries (1 for the categories and 10 for both the hyponyms and enumeration patterns) per term are performed sequentially, the method returns the results within a minute.

[query] [keyword]	[query] [keyword]
[query] en [keyword]	[keyword] en [query]
[keyword] en [query]	[query] en [keyword]
[keyword] [query]	[keyword] [query]
[keyword] zoals [query]	[query] of [keyword]
[query] en andere [keyword]	[query] en de [keyword]
[keyword] als [query]	[query] de [keyword]
[keyword] of [query]	[keyword] of [query]
[query] of [keyword]	[query] in [keyword]
[keyword] van [query]	[query] van [keyword]

Table 2. The 10 hyponym patterns (l.) and enumeration patterns (r.) used for GTAA.

9 Experimental Results

In this section we present experiments with two thesauri. In Section 9.1 we evaluate the performance when mapping terms to the Dutch GTAA for the audiovisual archives, where in Section 9.2 we use the United States NALT agricultural thesaurus.

To be of assistance, the relevant terms within the results of the method should be observable at a single glance. We therefore not only analyze the performance of the *best of three* term, but also the precision of the three individual methods and the average ranking of the terms in the benchmark set.

9.1 Experiments with the GTAA

We performed two experiments with the GTAA. In the first experiment, we map a set of 'expired terms' to the thesaurus, where in the second we use a 'leave-one-out' strategy to evaluate the recall and precision of the method. That is, we remove a term from the thesaurus and map this term.

Mapping expired terms to the GTAA. As novel items are constantly added to the archive of S & V, the GTAA is updated from time to time as well. A major problem is replacement of 'expired terms' with terms within the latest version of the GTAA.

In this experiment we discuss the applicability of our method to resolve this problem. As a benchmark set, S & V provided us a list of 78 pairs of such expired terms and the terms within the (current) GTAA to replace them.

We have automatically learned the patterns [15] for the hyponym and enumeration relations by selecting all terms in the thesaurus starting with a – e and their BT or RT respectively. For each of the two methods, we use the 10 patterns that are found to be most effective (Table 2). We use `[query]` as a placeholder for the term to be queried (thus outside the thesaurus) and `[keyword]` denotes the place in the snippets where we search for terms within the thesaurus. It is notable that there is an overlap between the patterns for the two relations. The patterns *zoals* and *en andere* are Dutch translations of the patterns first identified by Hearst [7] and translated by IJzereef in [24].

	USING CATEGORIES		WITHOUT CATEGORIES	
	correct	1 click away	correct	1 click away
best hyponym patterns	12	22	13	24
best enumeration patterns	4	16	2	13
best lexical	7	15	6	11
best of three	13	24	14	21
set of winners	18	41	20	41

Table 3. Performance for the 78 expired terms.

	USING CATEGORIES		WITHOUT CATEGORIES	
	correct	1 click away	correct	1 click away
recall with hyponym patterns	46	70	46	70
average ranking	9.45	7.72	14.84	8.14
recall with enumeration patterns	16	44	16	44
average ranking	7.31	3.47	7.43	3.63

Table 4. Recall and average rank for the 78 expired terms.

The results for the test with the expired terms can be found in Tables 3 and 4. Table 3 shows that the *best of three* provides the correct term in 13 cases, while the highest scored terms with the hyponym, enumeration and lexical methods are correct in only 12, 4, and 7 cases respectively. However, if we analyze the term selected as the *best of three*, then 24 out of 78 are directly linked to the term in the benchmark set by either the US, BT, NT or RT relation. Hence, the user can easily navigate to the best candidate. For 15 terms, no *best of three* could be identified.

We have also analyzed the recall of the terms that are *1 click away* from the benchmark term. For example, given the term *bioscooppersoneel* (see Table 1), the terms *personeel*, *werknemers*, *filmoperateurs*, *bioscopen*, *film* and *explicateurs* are all linked to this term. If the method select either one of these terms, the user can navigate in one step to the term *bioscooppersoneel*. The *set of winners* contains such a term in 41 out of the 78 cases.

The column at the righthand side of Table 3 shows that the performance of the lexical method improves when we take the category information into account. The results of the two search engine-based methods do not improve when using the category information.

Table 4 gives the number of terms and their average rank found by the web-based methods. The recall is best for the hyponym patterns. However, the average rank of both the benchmark term and of the terms with distance 1 is better using the enumeration patterns. Here we see that the use of category information has a positive effect on the ranking of the method using the hyponym patterns.

When analyzing the results of the methods, we encounter numerous terms found that are intuitively correct. In Table 5 we give a number of exipered terms, the *best of three* alternative found and the benchmark as given by S&V.

term	*best of three*	benchmark
tabaksplanten (tobacco plants)	*planten* (plants)	*tabak* (tobacco)
tabakswinkels (tobacco shops)	*winkels* (shops)	*detailhandel* (retail trade)
tegelzetters (tilers)	*bouwvakkers* (construction workers)	*ambachten* (crafts)
titanium	*metalen* (metals)	*chemische elementen* (chemical elements)
toxicologie (toxicology)	*geneeskunde* (medical science)	*vergiftigingen* (poisonings)
troepen (troops)	*militairen* (soldiers)	*krijgsmacht* (armed forces)
tweeverdieners (two-earner family)	*gezinnen* (families)	*inkomens* (incomes)

Table 5. Example terms with their 'best of three' and benchmark mapping.

	correct		correct
best lexical	138	recall with hyponym patterns:	343
best hyponym patterns	71	average ranking:	13.66
best enumeration patterns	87		
best of three	159	recall with enumeration patterns:	146
set of winners	239	average ranking:	2.12

Table 6. Performance, recall and average rank for the 573 GTAA terms

Leave one out. In this second experiment with the GTAA, we use the thesaurus itself as a benchmark set. We proceed as follows. We select a term t within the GTAA that has a broader term b. We then remove t from the thesaurus and use this thesaurus $T \setminus t$ as a reference. The task is now to find the mapping of the term outside the thesaurus (i.e. t) to b.

We use the same patterns as in the previous experiment. For fairness, we therefore will only evaluate with terms in the thesaurus starting with the letters *f* to *z* that have a broader term. This resulted in a test with 573 terms. When a term has multiple broader terms, in the evaluations we focus on the best scoring one.

The results for these tests are given in Table 6. It shows that in 239 out of the 573 terms (i.e. 42%) the correct term is among the set of winners (of size at most 4). In 92 cases the correct term was within the set of winners, but not found with the lexical approach. Table 6 shows again that the recall using the hyponym patterns is larger, but the ranking of the enumeration-based method is more precise. The lower recall using the enumerations can be explained by the structure of the thesaurus. As the GTAA is quite flat, for many terms found within the snippets no broader term is defined.

As an example, in Table 7 we give the best scoring output for the query term *fietspaden* (bicycle tracks). The broader term in the thesaurus is *infrastructuur* (infrastructure).

Given the difficulty of the tasks, and the fact that the mappings chosen in the benchmarks are sometimes debatable, we consider the results of the experiments convincing.

best of three:	paden	
lexical:	paden	*paths/tracks*
hyp. pat:	wegen	*roads*
hyp. pat:	trottoirs	*sidewalks*
hyp. pat:	paden	
hyp. pat:	meren	*lakes*
hyp. pat:	padden	*toads*
enum pat:	infrastructuur	*infrastructure*
enum pat:	paden	
enum pat:	openbare voorzieningen	*public services*
enum pat:	wegen	
enum pat:	beroepen	*professions*

Table 7. Best scoring output for *fietspaden* (bicycle tracks).

As the correct answer is present in the majority of the cases (as the hyponym pattern method found 343 out of 573 correct mappings), the method can be of value of an assistant for those working with the GTAA or the catalog of S & V. The experiment with the expired term showed that the determination of the categories improves the performance of the lexical approach. The results suggest that this preliminary step can be omitted for the other two approaches.

9.2 Experiment with NALT

To illustrate that the methods used are suited for English as well, we perform the last experiment with the the Agricultural Thesaurus (NALT) of the US National Agricultural Library.

The NALT contains Latin names for animals and plants, names for molecules and bacteria, but also product names such as *Brie Cheese*, *champagne* and *fish steaks*.

[query] [keyword]	[query] [keyword]
[query] and [keyword]	[query] and [keyword]
[query] are [keyword]	[keyword] and [query]
[query] or [keyword]	[query] or [keyword]
[keyword] and [query]	[query] are [keyword]
[query] the [keyword]	[query] such as [keyword]
[query] and other [keyword]	[query] of [keyword]
[keyword] such as [query]	[keyword] or [query]
[keyword] including [query]	[query] as [keyword]
[keyword] or [query]	[query] for [keyword]

Table 8. The 10 hyponym (l.) and enumeration (r.) patterns used for NALT.

	correct		correct
best lexical	169	recall with hyponym patterns:	468
best hyponym patterns	10	average ranking:	41.02
best enumeration patterns	100		
best of three	192	recall with enumeration patterns:	301
set of winners	286	average ranking:	8.39

Table 9. Performance for the 3321 NALT terms

We learn the patterns using the terms starting with the letter *a*. The patterns found for NALT are given in Table 8. As the NALT does not categorize the terms, we omit the step as described in Section 4. As the NALT consists of 68581 terms, we also leave out the collection of the number of search engine hits for each thesaurus term, as this would require 14 days using the *Yahoo!* API.

We test the performance of the methods on the terms within the 3321 NALT starting with the letters *b* to *z* that have a broader term. The results are given in Table 9. As an illustration Table 10 gives the output for the term *dietary cation anion difference*, where *feed formulation* is its broader term in the NALT. The right mapping for the (more common term) *bitterness* indeed was found (Table 11).

For the hyponym-pattern based approach, typically a long list of terms is found that all co-occur once with the queried term. As no frequency information is available, the ranking is just alphabetic. Using the enumeration method however, less terms are found. Moreover, as multiple hyponyms contribute to the score of their hypernym, the scores for the terms found tend to differ. Hence, although the number of correct mappings found with the enumeration patterns is smaller, the ranking of the method is much more reliable than the hyponym-based method.

hyp. pat:	ammonia	hyp. pat:	milk fever
hyp. pat:	anions	hyp. pat:	placenta
hyp. pat:	buffering capacity	hyp. pat:	retained placenta
hyp. pat:	fever	hyp. pat:	salts
hyp. pat:	literature	hyp. pat:	species differences
hyp. pat:	milk	hyp. pat:	urea
		hyp. pat:	viscosity
enum pat:	periparturient diseases and disorders		
enum pat:	pregnancy complications		

Table 10. The output for *dietary cation anion difference*.

It immediately shows that the results for NALT are far more modest. However, given the nature of the NALT and the fact that we did not correct for the frequencies of the terms, we consider the results as a proof of concept that this method is also applicable to another domain and language.

best of three:	flavor	
	...	emum pat: flavor
hyp. pat:	face	enum pat: ketones
hyp. pat:	families	enum pat: thermodynamics
hyp. pat:	fear	enum pat: physics
hyp. pat:	flavor	enum pat: light
hyp. pat:	food choices	enum pat: grapes
hyp. pat:	garlic	
	...	

Table 11. Part of the output for *bitterness*; 116 alternatives with the same score were found using the hyponym patterns.

10 Conclusions and Future work

We have developed an algorithm to assist people to find alternative terms within a thesaurus for a given query term.

The algorithm developed combines three approaches to map a term to a term within a thesaurus. We use both texts found with *Yahoo!* as well as a simple lexical matching technique to make the mapping. The algorithm is constructed independently from the content of the thesaurus and can easily be mapped to another language. The combination of independent methods lead to considerably better performances than any of the individual methods.

The method can facilitate searching a catalog with the use of index terms, since the algorithm can present a small number of alternatives thesaurus terms for a given term. This reduces the number of alternatives from the 5,000 GTAA terms to only a handful. The experiments with the GTAA show that the algorithm indeed can be usable as an assistant to find the right terms within the thesaurus.

In future work, we plan to test the method using queries collected from users of the web site of S & V. If common search terms are linked to the GTAA, the quality of the search results can improve. Apart from testing the method with terms given by the general public, we plan to let the method be evaluated by professional users as part of the thesaurus browser [27].

In our tests, some parameters, such as the number of patterns and the scores for the categories are chosen somewhat arbitrarily. In future work, optimal settings need to be determined. The use of the categories improved the performance of the lexical methods. Currently, it is not evident whether this preliminary step is needed for the other two methods.

11 Acknowledgements

We cordially thank Alma Wolthuis and Vincent Huis in ’t Veld from the Netherlands Institute for Sound & Vision for valuable discussions and providing the benchmark set as used in Section 9.1.

This research is carried out in the context of the Dutch BSIK MultimediaN project.

References

1. Shvaiko, P., Euzenat, J., Noy, N., Stuckenschmidt, H., Benjamins, R., Uschold, M., eds.: Proceedings of the ISWC'06 International Workshop on Ontology Matching. CEUR-WS Vol-225 (2006) http://www.om2006.ontologymatching.org/.
2. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. In: Journal on Data Semantics IV. Volume 3730 of Lecture Notes in Computer Science., Heidelberg, Germany, Springer (2005) 146 – 171
3. Aleksovski, Z., Klein, M.: Ontology mapping using background knowledge. In: Proceedings of Third International Conference on Knowledge Capture (K-CAP 2005), Banff, Canada (2005)
4. Meilicke, C., Stuckenschmidt, H., Tamilin, A.: Repairing ontology mappings. In: Proceedings of the 22nd Conference on Artificial Intelligence (AAAI-07), Vancouver, Canada (2007)
5. Gligorov, R., Aleksovski, Z., ten Kate, W., van Harmelen, F.: Using Google Distance to weight approximate ontology matches. In: Proceedings of the 16th international conference on World Wide Web (WWW2007), Banff, Canada (2007)
6. van Hage, W.R., Kolb, H., Schreiber, G.: A method for learning part-whole relations. In: Proceedings of the 5th International Semantic Web Conference (ISWC 2006). Volume 4273 of LNCS., Athens, GA, Springer (2006) 723 – 736
7. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics, Nantes, France (1992) 539 – 545
8. Hearst, M.: Automated discovery of wordnet relations. In Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)
9. Cimiano, P., Staab, S.: Learning by Googling. SIGKDD Explorations Newsletter **6**(2) (2004) 24 – 33
10. Etzioni, O., Cafarella, M.J., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence **165**(1) (2005) 91 – 134
11. Sumida, A., Torisawa, K., Shinzato, K.: Concept-instance relation extraction from simple noun sequences using a full-text search engine. In: Proceedings of the ISWC 2006 workshop on Web Content Mining with Human Language Technologies (WebConMine), Athens, GA (2006)
12. McDowell, L., Cafarella, M.J.: Ontology-driven information extraction with ontosyphon. In: Proceedings of the 5th International Semantic Web Conference (ISWC 2006). Volume 4273 of LNCS., Athens, GA, Springer (2006) 428 – 444
13. Tjong Kim Sang, E., Hofmann, K.: Automatic extraction of dutch hypernym-hyponym pairs. In: Proceedings of Computational Linguistics in the Netherlands (CLIN-17), Leuven, Belgium (2007)
14. Ravichandran, D., Hovy, E.: Learning surface text patterns for a question answering system. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, PA (2002) 41 – 47
15. Geleijnse, G., Korst, J.: Learning effective surface text patterns for information extraction. In: Proceedings of the EACL 2006 workshop on Adaptive Text Extraction and Mining (ATEM 2006), Trento, Italy (2006) 1 – 8
16. de Boer, V., van Someren, M., Wielinga, B.J.: A redundancy-based method for the extraction of relation instances from the web. International Journal of Human-Computer Studies **65**(9) (2007) 816–831
17. de Boer, V., van Someren, M., Wielinga, B.: Extracting art style periods from the web. In: Proceedings of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, Budva, Montenegro (2006)

18. Cilibrasi, R., Vitanyi, P.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Management **19**(3) (2007) 370 – 383
19. Geleijnse, G., Korst, J.: Creating a Dead Poets Society: Extracting a social network of historical persons from the web. In: Proceedings of the Sixth International Semantic Web Conference and the Second Asian Semantic Web Conference (ISWC + ASWC 2007). Volume 4825 of LNCS., Busan, Korea, Springer (2007)
20. Malaisé, V., Isaac, A., Gazendam, L., Brugman, H.: Anchoring dutch cultural heritage thesauri to wordnet: two case studies. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), Prague, Czech Republic (2007) 57 – 64
21. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)
22. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts (1999)
23. Geleijnse, G., Korst, J., de Boer, V.: Instance classification using co-occurrences on the web. In: Proceedings of the ISWC 2006 workshop on Web Content Mining with Human Language Technologies (WebConMine), Athens, GA (2006) http://orestes.ii.uam.es/workshop/3.pdf.
24. IJzereef, L.: Automatische extractie van hyponiemrelaties uit grote textcorpora. Master's thesis, Rijksuniversiteit Groningen, Groningen, the Netherlands (2004)
25. Snow, R., Jurafsky, D., Ng, A.Y.: Semantic taxonomy induction from heterogenous evidence. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL (COLING/ACL 2006), Sydney, Australia (2006) 801–808
26. Snow, R., Jurafsky, D., Ng, A.Y.: Learning syntactic patterns for automatic hypernym discovery. In: Advances in Neural Information Processing Systems 17, Cambridge, MA, MIT Press (2005) 1297–1304
27. Malaisé, V., Aroyo, L., Brugman, H., Gazendam, L., de Jong, A., Negru, C., Schreiber, G.: Evaluating a thesaurus browser for an audio-visual archive. In: Managing Knowledge in a World of Networks, 15th International Conference (EKAW 2006). Volume 4248 of Lecture Notes in Computer Science., Podebrady, Czech Republic, Springer (2006) 272–286