SemSearch 2008, CEUR Workshop Proceedings, ISSN 1613-0073, online at CEUR-WS.org/Vol-334/

The Interaction Between Automatic Annotation and Query Expansion: a retrieval experiment on a large cultural heritage archive

Véronique Malaisé[1], Laura Hollink[1], and Luit Gazendam[2]

[1] Department of Computer Science
Vrije Universiteit Amsterdam
de Boelelaan 1081 HV
The Netherlands
[2] Telematica Instituut
Brouwerijstraat 1
7523 XC Enschede
The Netherlands

Abstract. Improving a search system for large audiovisual archives can be done in two ways: by enriching the annotations, or by enriching the query mechanism. Both operations possibly benefit from a preliminary terminological enrichment of the controlled vocabulary in use, *i.e.* the thesaurus. In this paper we report on a four-parts experiment in which we evaluate the benefits and drawbacks of both aspects: the added value and pitfalls of automatically generated semantic annotations over classically (*i.e.* manually) assigned keywords and the added value and pitfalls of query expansion over pure keyword matching technique; we then investigate the combination of these operations in the following setup: we create the baseline for our experiments by querying a set of documents annotated by cataloguers with keywords from the thesaurus. We then apply the same querying process on a set of annotations automatically generated from textual resources related to the documents. Thirdly, we apply a querying process enhanced with query expansion functionalities to the first set of manually annotated documents. Finally, we apply the query expansion mechanism on the automatically generated annotations. The results give insight into the interaction between the two approaches.

1 Introduction

Enhancing the search results in large archives is a concern shared by many cultural heritage institutions. The improvement can come from two directions: enhancing the annotations or enhancing the search mechanism. Both directions are active research area's. In this paper we explore the interaction between those two approaches.

Enhancing the annotations can, for example, be done by facilitating manual creation of semantic annotations as in [10] or [4]. As manual annotation due to time constraints inherently leads to a relatively low number of keywords per

document, it can be complemented or even replaced by (semi-)automatically created annotations. In [13], for example, a tool is introduced for semi-automatic semantic annotation, extracted from text resources. Automatically generated annotations, however, seldom reach the quality level of manual annotations.

Another way to enhancing the search mechanism is query expansion: retrieval of not only documents that match the query concept, but also documents that are annotated with concepts that are *related* to the query. Ontology based query expansion is studied, for example, by [2]. The added value of query expansion in a cultural heritage archive has already been shown in [5]. However, the question remains what is the effect of query expansion in the context of automatic annotation? Is query expansion still beneficial when applied to lower-quality automatic annotations? And is it still necessary if a larger number of annotations is generated?

To answer these questions, we perform a study consisting of four experiments:

1. First, we compute a baseline by querying a corpus of hand-made metadata.
2. Second, we query the automatically generated annotations of the same corpus.
3. Third, we query the hand-made metadata using query expansion.
4. Fourth, we query the automatically generated annotations using query expansion.

The experiments that we present in this paper were conducted in collaboration with and on data from the Netherlands Institute for Sound and Vision, the Dutch national Audiovisual Archives. Our use-case consisting of audiovisual documents, we could have taken into account yet another field of research: the extraction of semantic keywords based on the video stream's low level features. As stated in [16], this technology is not really mature yet, and besides no detectors exist so far for the 3800 terms of the thesaurus we are interested in. Usually, the detectors are of hundreds of different types at most, and perform best on one given corpus of documents. For all these reasons, we took only into account so far the extraction based on textual descriptions of the audiovisual programs: extraction of keywords from textual resources gives good results. We did not take into account the transcripts form the videos either because of the numerous errors that these transcriptions contain: no NLP tool performs at an optimal level with syntactically incorrect sentences. Teletext and other resources will be used as input for our process at a later stage but as a first set of experiments we consider textual descriptions at a higher level of abstraction. This is the level that best suited our needs. Indeed, at Sound and Vision, the archived TV programs' *core topics* are described manually by cataloguers and annotated with keywords selected from a thesaurus, the GTAA. Our task is to extract keywords that describe as globally as possible the program's content.

The GTAA thesaurus is subsequently used for searching the archives. Its hierarchical structure is weak. As both query expansion and our automatic annotation mechanism rely on the structure of the thesaurus, we enriched the thesaurus with additional relations between its concepts.

In the remainder of this paper, we first describe the background on which the current paper is based: section 2 describes previous work on conversion of the thesaurus to SKOS, automatic semantic annotation, thesaurus enrichment and query expansion. Section 3 is dedicated to the description of the four experiments and their results. We conclude and propose future work in section 4.

2 Background

2.1 The GTAA thesaurus and its conversion to SKOS

The thesaurus that is used at Sound and Vision for the annotation and retrieval of TV programs stored in the archives is called the GTAA, a Dutch acronym for "Common Thesaurus [for] Audiovisual Archives". It is a faceted thesaurus, in Dutch, and each facet corresponds to at least one field in the document's description scheme. The topic(s) of the TV program is(are) described by terms from the Subject facet, which contains about 3800 Terms and 2000 additional variants of these terms such as so-called Nonpreferred Terms, which are not meant to be used for indexing but which aid in locating the right term. For example *posters* is a Nonpreferred Term that points to the term *affiches*, which is the right term to be used for indexing programs about posters, and is the only term that will enable a user to retrieve these documents. The Subject facet is organised according to hierarchical relationships (Broader Term/Narrower Term, between a term and its more general/specific notion) and associative relationships (Related Terms, such as *schepen* and *scheepvaart*, Dutch for respectively *ships* and *navigation(ship traffic)*). Besides these relationships defined in the ISO and NISO standards, the terms from the Subject facet are also grouped into a set of "topic" categories, like Philosophy, Economy, etc.

In order to use these relationships either in automatic annotation or query expansion processes, we converted the Subject facet to an RDF representation and modeled the relationships as SKOS triples [15]. For details about the conversion see [17].

2.2 Automatic semantic annotation

In the CHOICE project, we are using the GATE platform [6] for automatic generation of annotations from texts that are related to the TV programs. Other platforms and tool suits exist for generating ontology-based manual, semi-automatic or automatic annotations, like [13], but we chose GATE because we could use our own thesaurus as knowledge resource and tune the platform to our own needs. The idea that we are pursuing is to help cataloguers in their daily work with semi-automatic support. For this purpose, we have co-developed a plug-in called Apolda[3], which takes an ontology and a text as input, and returns an annotated text. The annotations refer to ontology URI (unique identifiers of concepts) and are based on the strings or *labels* that represent the ontology's

[3] Downloadable at the URL:`http://apolda.sourceforge.net/`

concepts for human readers. What we take at first for labels, in our case, are the Terms and Nonpreferred Terms of the GTAA: when they are matched in the text, an annotation is created, specifying the URI of the concept they refer to in the RDF version of the thesaurus. For example, a text containing both the words *posters* and *affiches* gets twice the annotation *GTAA_Subject_Posters*, their common URI. The texts we are using are called *context documents*, and describe the content of TV programs that will be or are stored in the archives: they are online TV guides or broadcaster's Websites for example. Besides the information already present in the thesaurus, we also computed the singular form of the Terms and Nonpreferred Terms based on the Celex lexicon [1], in order to get a better set of possible annotations. The possible annotations are meant as suggestions for annotating the TV programs the texts refer to.

The generated annotations contain sometimes long lists and/or errors due to the ambiguity of terms taken out of their context. In order to solve these two problems[4], we have developed a ranking algorithm. It is based on the structure of the thesaurus and a weighting system to compute the relative importance of the Terms matched in a given text. This algorithm is detailed below.

The semantic annotation pipeline The list of annotations that is extracted by Apolda along with their number of occurrences per textare fed to the CARROT algorithm. CARROT ranks highest the annotations that have direct and indirect thesaurus relationships to other annotations found for the same document, then the Terms that are connected to this group, then the annotations that have only indirect relationships to others, and finally then the rest.

In each of the aforementioned groups (annotations with direct and indirect, only indirect and no relationships to others at all), the annotations are further ordered based on a measure of their weight and their alphabetical order. The weighting of Terms' occurrences that we have experimented so far were pure occurrences counting and tf.idf weighting. For the experiments described in this paper, we also reduced the list of suggestion by taking into account only the first N ones, N being defined as the value of the square root of the list's length. We chose this value based on empirical tests: on average, only the part of the list that we kept are relevant annotation suggestions, the bottom of the list being filled mostly with noise.

For enhancing the search in the archives, a query expansion mechanism was developed in the context of the MUNCH project, aiming at multi-modal search in audiovisual archives.

[4] Having long lists of keywords extracted from texts is seen as a negative point because these lists are made to be shown to cataloguers, in order to speed up and ease their annotation process: showing them lists of more than hundred Terms is not an optimal solution in that respect, given the fact that their rules teach them to use as few of them as possible.

2.3 Query expansion

Like the semantic annotation, the query expansion mechanism is also based on the thesaurus structure. Thesaurus based query expansion requires a richly structured thesaurus. In previous experiments [11], we have show how we could use an anchoring of the GTAA to WordNet to add structure to the weakly structured GTAA. Wordnet is a terminological resource developed at the Princeton University [7], freely available from the Princeton website[5]. In addition, W3C has released a RDF/OWL representation of WordNet version 2.0[6]. For our experiment we use this RDF/OWL version, as it allows us to use Semantic Web tools such as SeRQL to query the WordNet database. We present here briefly the anchoring method that we used and the number of additional relationships inferred back in the original thesaurus, along with the process to infer them. We then go into the details of our query expansion mechanism.

Anchoring GTAA to WordNet As the GTAA is in Dutch, we queried an online dictionary in order to retrieve translations for the terms, along with definitions. Our purpose was to follow the method of [14] and base our anchoring on the lexical overlap between Term's descriptions and WordNet's descriptions: the glosses. The definitions that matched with the WordNet glosses, which was the case for more than 90 % of them, corresponded exactly to WordNet glosses, so the anchoring process was eased.

In total, 1,060 GTAA terms were anchored to WordNet. An evaluation of the correspondences suggests that the number of synsets that is aligned with a particular GTAA term is not an indication of the quality of the match; GTAA terms that are matched to six synsets are equally well matched as GTAA terms that are matched to only one synset.

Inferring additional relations in the GTAA We used the anchoring to WordNet to infer new relations within the GTAA. Using SeRQL [3] queries we related pairs of GTAA subject terms that were not previously related. 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 though either a direct (i.e. one-step) relation without any intermediate synsets or an indirect (i.e. two-step) relation with one intermediate synset. The latter situation is shown in Figure 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 [12] demonstrated that other types of WordNet relations do not improve

[5] http://wordnet.princeton.edu/
[6] http://www.w3.org/TR/wordnet-rdf/

retrieval results when used for query expansion. 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 [12, 9] showed that changing direction, especially in the hyponym/hypernym hierarchy, decreases semantic similarity significantly. For example, w_1 holonym of w_i hyponym of w_2 is not included. At present, all anchoring relations are utilized, also the ones that relate a GTAA term to multiple WordNet terms.

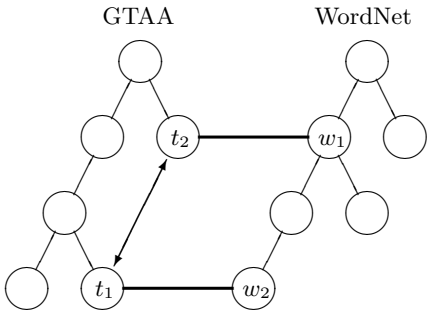

Fig. 1. Using the anchoring to WordNet to infer relations within the GTAA.

A total of 904 pairs of GTAA terms was newly related: 467 with one step between WordNet synsets w_1 and w_2 and 435 with 2 steps between w_1 and w_2. An inspection of the inferred relations reveals that 90 % of the one-step relations were derived from hyponym relations and only 10% from meronym relations. The two-step relations were for 72 % based on sequences of two hyponym relations, for 26 % on combinations of hyponym and meronym and only for 3 % on sequences of two meronym relations.

An informal manual inspection of a portion of the new relations revealed that only very few seem wrong. Based on the original GTAA and the newly inferred relationships, we implemented a query expansion mechanism dedicated to Sound and Vision, but its general mechanism can be applied to any archive using a thesaurus for annotating their data.

The query expansion mechanism Query expansion was done by simply adding concepts to the query that are a fixed number of steps away from the original query concept. All relations were used to walk through the thesaurus: broader, narrower, related, but also the relations inferred from the links to WordNet.

We experimented with expansion to concepts that were only one step away from the query, and with expansion to concepts up to two steps away. As the GTAA has a shallow structure, expanding a query with concepts that are more than two steps away leads too often to concepts that are in an unrelated part of the hierarchy.

2.4 Related work

As we did experiments on both types of methods for enhancing the search process in large archives, we wanted to test how these techniques would interact and what their combination would bring. In the litterature, see [18] for example, either one or the other of the aspects are investigated, namely either improvement based on semantic annotation or on query expansion. We chose to analyze their combination and ran a set of four experiments, described in more details in the following section.

3 Four Experiments

3.1 Material: queries, test corpus and gold standard

In order to be as close as possible from a real-life need, we selected a set of queries from one week of query logs collected at Sound and Vision. We selected the top 44 in the list of most frequently asked keywords, in the keyword search field of the query interface, and stoped the selection with the group of keywords that had only two occurrences in the query log.

The list of the top 44 keywords is: Geschiedenis (history), Kabinetsformaties (forming of parliament), Parlementaire debatten (parliamentary debates), Politici (politicians), Politiek (politics), Politieke partijen (political parties), Politieke programma's (political programmes), Verkiezingen (elections), Verkiezingscampagnes (election campaigns), Gemeenteraden (municipal councils), Asielzoekers (asylum seekers), Islam (islam), Leger (army), Mobilisatie (mobilisation(of army)), Atoombommen (nuclear bombs), Bombardementen (bombardments), Explosies (explosions), Gevaarlijke stoffen (dangerous substances), Gewonden (wounded), Eerste hulp (first aid), Geneesmiddelen (medications), Euthanasie (euthanasia), Dementie (dementia), Broeikaseffect (greenhouse effect), File's (traffic-jams), Snelwegen (highways), Spoorwegongevallen (railway accidents), Autobussen (busses), Alcohol (alcohol), Cafe's (cafe's), Fabrieken (factories), CAO's (collective work agreements), Vulkaanuitbarstingen (vulcano eruptions), Woestijnen (deserts), ‚Zonsondergangen (sunsets), Voetbal (soccer), Zwembaden (swimming pools), Schaatsen (ice skating), Kaartspelen (cardgames), Kermissen (village fairs), Mode (fashion), Opvoeding (education), Dierenhandel (animal trade), Grachten (canals).

These 44 queries are matched against a textual corpus that we had built for previous experiment according to the following rationales:

- The corpus is focused on TV program's description made manually by cataloguers and stored in the previous system for managing the archives at Sound and Vision: Avail. We therefore call these manual catalogue entries "Avail documents"[7];

[7] These can be accessed online at `http://www.beeldengeluid.nl/collecties_zoek_en_vind_tvfilm.jsp`.

– We only selected descriptions of programs which were part of a collection called Academia [8];
– We only selected descriptions of programs for which we could find open accessible context documents: textual descriptions of the TV program's content on broadcaster's Websites or TV-guides, for example;
– We narrowed our selection to documentary programs.

The choice of limiting ourselves to documents related to the academia collection and to documentaires is explained by the fact that, on the one hand, the Academia collection has been cleared from intellectual property rights by Sound and Vision in order to create an open accessible collection for educational and research purposes. Although we do not use this primary audiovisual content in this research, we decided that it would be wise restrict our corpus selection to documents with open accessible AV material.

On the other hand, we narrowed down our selection to documentary programs for multiple reasons: (1) they usually had accessible context information such as web sites, even though some programs could be as old as 7 years. For news items, sport programs or actualities this is not the case. This made the manual selection much more efficient. (2) the information described in their context documents is usually quite extensive. Because we want to gain insight into the process of annotating via context documents, we wanted to have as few content-wise difference with the actual AV document content.

For all the web sites, these textual resources were selected and copied manually. Table 1 details the composition of the corpus.

Series name	Program topic	nb of programs
andere tijden	history	93
beeldenstorm	art	68
de donderdag documentarie	humanities	6
de nieuwe wereld	informative	5
dokument	humanities	6
dokwerk	history or politics	57
Jota!	science	10
Nieuw economisch peil	economy	10
werelden	social	3

Table 1. The composition of our corpus

3.2 Experiment one: the baseline

The baseline experiment consisted in evaluating how many of the Avail documents were annotated with one or more of the "Top 44" keywords. As the assessment of keywords was done by hand and as we evaluate queries consisting in only one keyword, if the keyword is present in the Avail metadata[8], we

[8] The keyword field of the metadata only, to be more specific.

consider that the document is relevant for that keyword. In order to have an idea about the recall, we computed an "estimated recall" by evaluating how many of the documents from the golden standard that we judged relevant to be annotated by one of these 44 keywords were retrieved (column "Estimated recall" from the "Manual Metadata" section in table 2). Our most successful keyword (*geschiedenis*, Dutch for *history*) retrieved 97 documents, but most of the keywords (14) did not retrieve any document in our test corpus. The section "Manual Metadata" of table 2 shows the number of documents retrieved per keyword and the estimated recall, based on our gold standard. The estimated recall is labelled as "Non relevant" (NR) if there were no documents annotated by this keyword in our manually established golden standard.

One first remark that we can derive from this table is the low values for estimated recall. It can be due to two reasons. Firstly, we evaluated whether a set of 44 queries was suitable for annotating documents, whereas the cataloguers have a larger choice: they can select any term from a set of 3800. Therefore the granularity level and the selection can be quite different (for example, they would probably choose *second world war* where we judged that *army* was relevant as a keyword). Secondly, some of the keywords, like *politicians* or *political parties*, can be replaced by a list of names corresponding to the people or parties mentioned in the TV programs. A cataloguer from Sound and Vision would choose this option, as it gives more precise information than the generic Subject keyword. As our experiment focuses only on Subject keywords, and not on the other parts of the metadata, and as there is not built-in relationship between names (of politicians or political parties) and their types in the thesaurus, we could not bridge this gap. But this problem is interesting to keep in mind for providing more relevant automatic semantic annotations in the future, by creating automatically this missing link.

3.3 Experiment two: keyword matching on automatic semantic annotations

After computing the baseline with the first experiment, we applied the same evaluation metrics to the annotations generated automatically by our semantic annotation pipeline: we counted the number of documents that were retrieved for each of the 44 queries, we estimated a recall measure based on the number of documents from our gold standard that were retrieved. We also computed the overlap between the documents that were retrieved based on manual annotation and documents retrieved based on annotations that were generated with the Apolda plugin. This is show in the column called 'overlap' in Table 2.

Queries based on the manually assigned annotations retrieved 142 documents, with an average recall of 22.3 %. Nine queries retrieved documents out of 26 possibilities in our manually established golden standard. The figures are not that good for the queries that were matched against the automatically generated keywords: only 57 documents were retrieved, with an average estimated recall of 9.6% and only 6 keywords out of the 26 possible retrieved documents. Here again, the explanation is twofold. On the first hand, our random sample of documents

	Manual Metadata		Automatic Metadata		Overlap
Query	retrieved	estimated recall	retrieved	estimated recall	
History	97	23/60=38.33	6	1/60=1.66	2
Forming of Parliament	0	NR	0	NR	NR
Parlementary debates	0	NR	0	NR	NR
Politicians	2	0/14=0	3	6/14=42.85	0
Politics	10	1/15=6.66	8	1/15=6.66	3
Political parties	2	NR	0	NR	0
Political programmes	0	0/1=0	0	0/1=0	NR
Elections	1	1/1=100	4	1/1=100	1
Election campaigns	3	1/1=100	0	0/1=0	0
Municipal councils	0	0/2=0	1	1/2=50	0
Asylum seekers	7	0/2=0	2	0/2=0	2
Islam	3	0/4=0	3	0/4=0	2
Army	1	1/7=14.28	9	2/7=28.57	1
Military mobilisation	0	0/1=0	0	0/1=0	NR
Nuclear bombs	1	NR	2	NR	1
Bombardments	2	0/2=0	1	0/2=0	1
Explosions	0	0/1=0	3	0/1=0	0
Dangerous substances	0	0/4=0	1	0/4=0	0
Wounded	1	0/5=0	1	1/5=20	0
First aid	0	NR	0	NR	NR
Medications	2	0/2=0	0	0/2=0	0
Euthanasia	0	NR	0	NR	NR
Dementia	0	0/1=0	0	0/1=0	NR
Greenhouse gas effect	0	NR	0	NR	NR
Traffic jams	0	NR	1	NR	0
Highways	0	0/2=0	1	0/2=0	0
Railway accidents	0	0/1=0	0	0/1=0	NR
Busses	1	NR	2	NR	0
Alcohol	0	NR	1	NR	0
Cafe's	0	NR	0	NR	NR
Factories	0	0/8=0	0	0/8=0	NR
Collective Work Agreement	0	0/3=0	1	0/3=0	0
Volcano eruption	0	NR	0	NR	NR
Deserts	1	1/1=100	0	0/1=0	0
Sunsets	0	NR	1	NR	0
Soccer	3	2/2=100	0	0/2=0	0
Swimming pools	0	NR	0	NR	NR
Ice skating	1	NR	2	NR	0
Card games	0	NR	0	NR	NR
Village fairs	0	0/1=0	0	0/1=0	NR
Fashion	1	1/1=100	0	0/1=0	0
Eduction	3	1/5=20	3	0/5=0	0
Animal trade	0	NR	0	NR	NR
Canals	0	NR	1	NR	0

Table 2. Retrieval results of experiments one and two: keyword search on manually made annotations and automatically generated annotations.

constituting the golden standard contained 97 documents describing the TV series Andere Tijden about history, and the whole collection is annotated by *history*. As all the documents deal with history, the word itself is seldom present in texts describing the content of the individual TV programs of the series, hence our automatic annotation pipeline could not achieve the recall that was obtained by querying on the manual metadata. Here again, this problem shows a point to keep in mind for improving our automatic annotation tool: we need to generate

also keywords that are relevant for the whole series of TV programs and not only for the individual ones.

An interesting point to notice, though, is that the Apolda-based annotations enables us to retrieve a document from the Art documentaries serie that was not annotated with *history* by cataloguers, but was judged relevant in our gold standard. Another possible explanation of the poor performance of the queries ran on Apolda annotations is the fact that they are quite generic, and the Top 44 queries extracted from the query logs are very specific. Thus, they are closer to what cataloguers do as manual annotation than to our automatically generated ones. This distance should be bridged by using a query expansion mechanism, option that we test in the next set of experiments.

Another thing that we can notice is that out of the total number of 199[9] retrieved documents, only 13 were overlapping between the results of the queries based on Avail or Apolda keywords. This number tends to suggest that the two approaches, rather than building one on the other, are complementary and should be run in parallel. A manual check of the retrieved documents that were part of the golden standard shows us that there is also a few overlap in terms of retrieved documents and successful queries, which reinforces our impression of complementary approaches.

3.4 Experiment three: query expansion on manual annotations

While in experiments one and two we retrieved documents based on an exact match between query and annotation concept, in experiments three and four we employ query expansion: we also retrieve documents that are annotated with concepts related to the query concept. We experiment with expansion to concepts that are one or two steps away from the query concept. The results are shown in table 3, agregating the results from experiments 3 and 4. The queries are ordered by decreasing number of hits.

In experiment three, query expansion is done on the manually created annotations. Using one-step expansion, this results in on average 7.6 documents per query. Two-step expansion retrieves four times as many documents: 28.2 on average. As expected, recall is higher than the recall in experiment 1 (37% for one-step and 58% for two-step expansion, compared to 22% in experiment 1), but precision is low (43% and 21% on average). With query expansion, documents are retrieved for 35 (one-step) or 38 (two-step) of the 44 queries. This is considerably more than in experiment 1, where documents were returned for only 19 queries.

3.5 Experiment four: query expansion on automatic semantic annotations

In experiment four, we apply query expansion to automatically generated annotations. One-step query expansion resulted in a mean of 8.6 retrieved documents,

[9] 142+57 documents, by summing up the total amount of the documents retrieved by the queries on the keywords either assigned manually or generated automatically.

two-step expansion in 40.3 documents. The combination of two-step query expansion with automatically generated annotation appears to lead to a strong increase in the number of retrieved docs. Precision is 0.29 for one-step and 0.11 for two-step expansion; recall is 0.30 and 0.48 respectively. A comparison of experiment two to the baseline showed that the Apolda annotations perform worse than the manually assigned annotations. A comparison of experiment three to experiment four paints a similar picture: both precision and recall of experiment four are lower than the query expansion results on manually created annotations in experiment three.

The results further show that where automatic annotations perform poorly when we search for an exact match with a query concept (experiment 2), they do lead to acceptable results when combined with query expansion (experiment 4). This combined strategy returns documents for 41 out of 44 queries.

The overlap between what is found using manual annotations and what is found using the automatically generated annotations is small. If expansion is limited to one step the overlap is 2.3 documents on average. Two-step expansion shows an overlap of 13.8 documents, which is relatively larger but still low. This suggests that it is worthwhile to add automatic annotations also in situations where good manual annotations are available.

The general table (table 3) give rise to some comments: theoretically, broadening the query expansion mechanism by taking into account Terms that ar at a distance 2 from the query Term could lead to one of the following outcomes:

- the query expansion heightens F score (The loss in precision is much lower than the gain in recall);
- the query expansion does not really influences Fscore (a loss in precision is compensated by a rise in recall);
- the query expansion lowers the F score (loss in precision is much larger than the gain in recall);

Interestingly enough, we see all three outcomes in our results. Therefore we cannot make a global conclusion about whether taking one only or the full two steps into account for query expansion is good or not in general, but we can see some properties of the Terms that would enable us to make a choice in some cases. For the Terms that have a high precision and low recall with one step of query expansion, like *education* or *collective work agreement*, one extra step gives a better recall without a big loss in precision. This heuristic holds for both Manual and Automatic metadata. For some Terms, we can observe the inverse: for example for *elections* or *election campaigns*, one step of query expansion already gives a low precision and 100% recall, for both Apolda and Avail. For these Terms, taking into account a second step only lowers the precision. For the third case, we cannot decide on a heuristic, as the F-measure is neither improved or jeopardised. The difference between the two first cases is strongly related to the structure of the thesaurus, which is not homogeneous: some Terms are in broad hierarchies (up to 7 levels down), whereas some Terms are not related to any other in the thesaurus. Thus, it is the results of the narrowest possible query

expansion that gives us the means to decide for the relevance of taking a broader one into account.

4 Conclusion and perspectives

We presented a set of four experiments in this paper, a baseline measurement and three possible ways to improve the retrieval results of this first baseline. One experiment involved automatic annotation and the two other experiments were based on query expansion mechanisms. It turned out that the automatic annotation setting performed worse than the baseline, when looking only at the numbers. But a qualitative look at the results showed us a very nice feature: the few overlap between the retrieved documents and the successful queries in the two settings make them quite complementary. Besides, one of the drawbacks of the automatic annotation is the genericity of the Terms extracted, which can be corrected by the query expansion mechanisms. The results of the fourth experiment confirm this hypothesis: the improvement of the automatic annotation-based setting was greater than the one based on manual annotations, with still a small overlap in the results. The complementarity of the two approaches is thus underlined, and could suggest to use them both in order to improve the search in large archives: adding automatic annotations to existing ones for a large archive could be a way of improving the accessibility of its content at low costs. The query expansion results improved the results, but also showed us the influence of the structure of the thesaurus in its performance: to get better performances by taking into account one or two steps of thesaurus relationships from a Term depends on the richness of the relationships network of that given Term. A two-times approach seems to be better suited to get the best possible results.

These experiments gave us some insights about improvements to add to our automatic annotation pipeline and query expansion mechanisms, and gave us interesting lines for future research: having a closer look at the influence of the relationships' network in the thesaurus and compensating for its non homogeneity in query expansion, using information provided by other metadata values (like the names of the people mentioned in the document) either for query expansion or semantic annotation.

Acknowledgements

This project was done in the context of the CHOICE and MUNCH projects, both part the NWO CATCH program. We would like to thank our colleagues from the Netherlands Institute for Sound and Vision who support us in our research.

References

[1] R. H. Baayen, R. Piepenbrock, and L. Gulikers. *The CELEX Lexical Database*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA:, (release 2) [cd-rom] edition, 1995.

[2] J. Bhogalb, A. Macfarlanea, and P. Smitha. A review of ontology based query expansion. *Information Processing & Management*, 42(4):866–886, July 2007.

[3] Jeen Broekstra and Arjohn Kampman. SeRQL: A second generation RDF query language. In *Proceedings of the SWAD-Europe Workshop on Semantic Web Storage and Retrieval*, pages 13–14, Amsterdam, The Netherlands, November 2003.

[4] F. Ciravegna and Y. Wilks. Designing Adaptive Information Extraction for the Semantic Web in Amilcare. In S. Handschuh and S. Staab, editors, *Annotation for the Semantic Web*. IOS Press, Amsterdam, 2003.

[5] Daniel Cunliffe, Carl Taylor, and Douglas Tudhope. Query-based navigation in semantically indexed hypermedia. In *HYPERTEXT '97: Proceedings of the eighth ACM conference on Hypertext*, pages 87–95, New York, NY, USA, 1997. ACM.

[6] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[7] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.

[8] Beeld & Geluid. academia collectie. http://www.academia.nl.

[9] Graeme Hirst and David St-Onge. *Lexical chains as representations of context for the detection and correction of malapropisms*, chapter 13, pages 305–332. The MIT Press, Cambridge, MA, USA, 1998.

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

[11] Laura Hollink, Véronique Malaisé, and A. Th. Schreiber. Enriching a thesaurus to improve retrieval of audiovisual material. Submitted for publication.

[12] Laura Hollink, Guus Schreiber, and Bob Wielinga. Patterns of semantic relations to improve image content search. *Journal of Web Semantics*, 5:195–203, 2007.

[13] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, December 2004.

[14] K. Knight and S. Luk. Building a large-scale knowledge base for machine translation. In *the AAAI-94 Conference*, 1994.

[15] Alistair Miles and Dan Brickley. SKOS core guide. W3C working draft, November 2005. Electronic document. Accessed February 2008. Available from: http://www.w3.org/TR/swbp-skos-core-guide/.

[16] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[17] Mark van Assem, Veronique Malaise, Alistair Miles, and Guus Schreiber. A method to convert thesauri to skos. In *Proceedings of the Third European Semantic Web Conference (ESWC'06)*, number 4011 in Lecture Notes in Computer Science, pages 95–109, Budva, Montenegro, June 2006.

[18] M. Volk, B. Ripplinger, S. Vintar, Paul Buitelaar, D. Raileanu, and B. Sacaleanu. Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 1/3(67):79–112, 2002.

	One-step Query Expanion							Two-step Query Expanion						
	Manual metadata			Automatic metadata			Overlap	Manual metadata			Automatic metadata			Overlap
Query	**retrieved**	**precision**	**estimated recall**	**retrieved**	**precision**	**estimated recall**		**retrieved**	**precision**	**estimated recall**	**retrieved**	**precision**	**estimated recall**	
Politics	17	0.75	3 / 15 = 0.2	23	0.5	4 / 15 = 0.27	9	50	0.42	5 / 15 = 0.33	59	0.35	6 / 15 = 0.4	31
Deserts	3	1	1 / 1 = 1	5	0	0 / 1 = 0	0	59	0.08	1 / 1 = 1	60	0	0 / 1 = 0	37
Education	7	1	1 / 5 = 0.2	11	0.5	1 / 5 = 0.2	3	23	0.67	4 / 5 = 0.8	47	0.43	3 / 5 = 0.6	13
Elections	13	0.33	1 / 1 = 1	18	0.25	1 / 1 = 1	6	22	0.2	1 / 1 = 1	39	0.09	1 / 1 = 1	12
Ice skating	3	NR	0 / 0 = NR	5	0	0 / 0 = NR	1	11	0	0 / 0 = NR	14	0	0 / 0 = NR	5
Army	12	0.67	2 / 7 = 0.29	23	1	4 / 7 = 0.57	10	35	0.5	4 / 7 = 0.57	59	0.4	4 / 7 = 0.57	23
History	96	1	23 / 32 = 0.72	5	1	2 / 32 = 0.06	5	106	0.88	23 / 32 = 0.72	25	0.88	7 / 32 = 0.22	12
Asylum seekers	22	0.4	2 / 2 = 1	7	NR	0 / 2 = 0	5	46	0.17	2 / 2 = 1	21	0	0 / 2 = 0	14
Nuclear bombs	1	NR	0 / 0 = NR	2	NR	0 / 0 = NR	1	7	NR	0 / 0 = NR	27	0	0 / 0 = NR	3
Medications	11	0	0 / 2 = 0	8	0	0 / 2 = 0	1	70	0	0 / 2 = 0	103	0.04	1 / 2 = 0.5	36
Islam	11	1	3 / 5 = 0.6	11	0.33	1 / 5 = 0.2	6	44	0.56	5 / 5 = 1	57	0.08	1 / 5 = 0.2	26
Busses	5	0	0 / 0 = NR	10	0	0 / 0 = NR	4	19	0	0 / 0 = NR	34	0	0 / 0 = NR	15
Fashion	10	0.5	1 / 1 = 1	2	NR	0 / 1 = 0	0	29	0.09	1 / 1 = 1	35	0	0 / 1 = 0	10
Politicians	16	0.25	1 / 9 = 0.11	27	0.5	4 / 9 = 0.44	6	68	0.33	6 / 9 = 0.67	114	0.28	8 / 9 = 0.89	46
Political parties	12	0	0 / 0 = NR	10	0	0 / 0 = NR	5	18	0	0 / 0 = NR	31	0	0 / 0 = NR	9
Factories	12	1	5 / 8 = 0.62	18	0.83	5 / 8 = 0.62	7	49	0.5	7 / 8 = 0.88	71	0.25	5 / 8 = 0.62	29
Bombardements	11	0	0 / 2 = 0	19	0.33	1 / 2 = 0.5	5	50	0.08	1 / 2 = 0.5	73	0.11	2 / 2 = 1	33
Wounded	5	NR	0 / 5 = 0	9	0.25	1 / 5 = 0.2	2	33	0	0 / 5 = 0	28	0.14	1 / 5 = 0.2	12
Soccer	3	1	2 / 3 = 0.67	0	NR	0 / 3 = 0	0	6	0.67	2 / 3 = 0.67	6	0	0 / 3 = 0	1
Election campaigns	5	0.5	1 / 1 = 1	6	1	1 / 1 = 1	2	24	0.2	1 / 1 = 1	28	0.17	1 / 1 = 1	11
Card games	1	0	0 / 0 = NR	2	NR	0 / 0 = NR	0	9	0	0 / 0 = NR	8	0	0 / 0 = NR	2
Swimming pools	2	NR	0 / 0 = NR	2	NR	0 / 0 = NR	1	52	0	0 / 0 = NR	98	0	0 / 0 = NR	28
Village fairs	1	NR	0 / 1 = 0	4	0	0 / 1 = 0	0	15	0	0 / 1 = 0	37	0	0 / 1 = 0	6
Military mobilisation	1	0	0 / 1 = 0	1	NR	0 / 1 = 0	0	20	0	0 / 1 = 0	28	0.17	1 / 1 = 1	9
Traffic jams	1	0	0 / 0 = NR	12	0	0 / 0 = NR	0	30	0	0 / 0 = NR	40	0	0 / 0 = NR	12
Collective Work Agreement	3	0.5	1 / 3 = 0.33	3	1	1 / 3 = 0.33	2	12	0.43	3 / 3 = 1	22	0.29	2 / 3 = 0.67	8
Animal trade	3	0	0 / 0 = NR	8	0	0 / 0 = NR	2	25	0	0 / 0 = NR	60	0	0 / 0 = NR	15
First aid	5	NR	0 / 0 = NR	6	0	0 / 0 = NR	0	32	0	0 / 0 = NR	35	0	0 / 0 = NR	17
Euthanasia	6	0	0 / 0 = NR	5	NR	0 / 0 = NR	1	25	0	0 / 0 = NR	26	0	0 / 0 = NR	8
Cafs	6	NR	0 / 0 = NR	16	0	0 / 0 = NR	4	63	0	0 / 0 = NR	104	0	0 / 0 = NR	36
Forming of Parliament	4	0	0 / 0 = NR	23	0	0 / 0 = NR	2	16	0	0 / 0 = NR	38	0	0 / 0 = NR	11
Political programmes	12	0	0 / 1 = 0	8	0	0 / 1 = 0	3	26	0.14	1 / 1 = 1	51	0.07	1 / 1 = 1	15
Canals	7	0	0 / 0 = NR	17	0	0 / 0 = NR	4	30	0	0 / 0 = NR	62	0	0 / 0 = NR	16
Explosions	3	1	1 / 2 = 0.5	11	0.17	1 / 2 = 0.5	2	7	1	1 / 2 = 0.5	23	0.14	1 / 2 = 0.5	2
Railways accidents	4	1	1 / 2 = 0.5	10	0.2	1 / 2 = 0.5	1	15	0.25	1 / 2 = 0.5	27	0.09	1 / 2 = 0.5	8
Sunsets	0	NR	0 / 0 = NR	1	NR	0 / 0 = NR	0	2	NR	0 / 0 = NR	11	0	0 / 0 = NR	0
Municipal councils	0	NR	0 / 2 = 0	3	1	2 / 2 = 1	0	11	0.5	1 / 2 = 0.5	15	0.5	2 / 2 = 1	3
Highways	0	NR	0 / 2 = 0	12	0.25	1 / 2 = 0.5	0	6	1	1 / 2 = 0.5	25	0.14	1 / 2 = 0.5	2
Parlementary debates	0	NR	0 / 0 = NR	10	0	0 / 0 = NR	0	25	0	0 / 0 = NR	39	0	0 / 0 = NR	9
Alcohol	0	NR	0 / 0 = NR	1	NR	0 / 0 = NR	0	13	0	0 / 0 = NR	16	0	0 / 0 = NR	5
Dementia	0	NR	0 / 1 = 0	0	NR	0 / 1 = 0	0	2	0	0 / 1 = 0	6	0	0 / 1 = 0	0
Greenhouse gas effect	0	NR	0 / 0 = NR	2	0	0 / 0 = NR	0	6	0	0 / 0 = NR	27	0	0 / 0 = NR	3
Dangerous substances	0	NR	0 / 4 = 0	1	NR	0 / 4 = 0	0	16	0	0 / 4 = 0	28	0.17	1 / 4 = 0.25	7
Volcano eruption	0	NR	0 / 0 = NR	0	NR	0 / 0 = NR	0	12	0	0 / 0 = NR	14	0	0 / 0 = NR	6

Table 3. Retrieval results of experiments two and three: query expansion on the manually made metadata and the automatically generated metadata.

SemSearch 2008, CEUR Workshop Proceedings, ISSN 1613-0073, online at CEUR-WS.org/Vol-334/