Automatic Annotation Suggestions for Audiovisual Archives: Evaluation Aspects

Luit Gazendam[1], Véronique Malaisé[2], Annemieke de Jong[3], Christian Wartena[1], Hennie Brugman[4], and Guus Schreiber[2]

[1] Telematica Instituut, Enschede, The Netherlands
[2] Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands
[3] Netherlands Institute for Sound and Vision, Hilversum, The Netherlands
[4] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Abstract. In the context of large and ever growing archives, generating annotation suggestions automatically from textual resources related to the documents to be archived is an interesting option in theory. It could save a lot of work in the time-consuming and expensive task of manual annotation and it could help cataloguers attain a higher inter annotator agreement. However, some questions arise in practice: what is the quality of the automatically produced annotations? How do they compare with manual annotations and with the requirements for annotation that were defined in the archive? If different from the manual annotations, are the automatic annotations wrong?
In the CHOICE project, partially hosted at the Netherlands Institute for Sound and Vision, the Dutch public archive for audiovisual broadcasts, we automatically generate annotation suggestions for cataloguers. In this paper, we define three types of evaluation of these annotation suggestions: (1) a classic and strict precision/recall measure expressing the overlap between automatically generated keywords and the manual annotations, (2) a loosened precision/recall measure for which semantically very similar annotations are also considered as relevant matches, (3) an in-use evaluation of the usefulness of manual versus automatic annotations in the context of Serendipitous Browsing. During serendipitous browsing the annotations (manual or automatic) are used to retrieve and visualize semantically related documents.

1 Context

The Netherlands Institute for Sound and Vision (henceforth S&V) is in charge of archiving publicly broadcasted TV and radio programs in the Netherlands. Two years ago the audiovisual production and archiving environment changed from analogue towards digital data. This effectively quadrupled the inflow of archival material and as such the amount of work for cataloguers. The two most important customer groups are: 1) professional users from the public broadcasters and 2) users from science and education. These typically have three kinds of user queries:

1. Known items queries: *e.g.* the eight o' clock news of 21-12-1976.
2. Subject queries: *e.g.* broadcasts with *ethnical minorities* as topic.
3. Shots and quotes: *e.g.* a fragment in which Barrack Obama says "Yes we can!"

S&V faces the challenge to create a durable continuous access to the daily increasing collections with the same number of cataloguers (40 people). The manual annotation is the bottleneck in the archiving process: it may take a cataloguer up to three times the length of a TV program to annotate it manually, depending on the genre (news item, game-show, documentary). During annotation, cataloguers often consult and use available contextual information such as TV-guide synopses, official TV programs web site texts and subtitles.

The annotation process follows strict guidelines. All catalogue descriptions conforms to a metadata scheme called iMMiX. The iMMiX metadata model is an adaptation for 'audiovisual' catalogue data of the FRBR data model [1] which has been developed in 1998, by the international federation of library associations (IFLA).

The iMMiX metadata model captures four important aspects of a broadcast:

1. information content (Who, what, when, where, why and how, includes keywords, organizations, locations)
2. audiovisual content (What can be seen or heard? Includes descriptions like *close-up*)
3. formal data, (*e.g.* intellectual property rights)
4. document management data (*e.g.* document ID)

Choices for some of the iMMiX fields (subject, location, persons etc.) are restricted to a controlled vocabulary named GTAA. GTAA is a Dutch acronym for "Common Thesaurus [for] Audiovisual Archives" and contains about 160 000 terms, organized in 6 facets. The GTAA subject facet contains 3800 keywords and 21000 relations between the keywords belonging to the ISO-2788 defined relationships of Broader Term, Narrower Term, Related Term and Use/Use for. It also contains linguistic information such as preferred textual representations of keywords and non-preferred representations. Each keyword averagely has 1 broader, 1 narrower and 3.5 related terms. Cataloguers are instructed to select keyword that describe the program as a whole, are specific and allow good retrieval.

1.1 Automatic Annotation Suggestions in the Choice-Project

Within this context, the CHOICE project investigates how to automatically suggest GTAA keywords to cataloguers during their annotation task. We assume that by applying Natural Language Processing and Semantic Web techniques to the contextual information (*e.g.* TV guide texts describing the broadcast),

[1] Functional Requirements for Bibliographical record, www.ifla.org/VII/s13/frbr/frbr.pdf

reasonable annotation suggestions can be generated. These suggestions are intended to increase a cataloguers working speed and consistency. Typical measures of inter-cataloguer consistency range from 13% to 77% (with an average of 44%) when a controlled vocabulary is used(Leininger 2000). The topology of disagreement shows that a portion of these differences are small semantic differences. This disagreement can be problematic when manual annotations serve as a gold standard for the evaluation of our automatic annotation suggestions. Nevertheless, the manual annotations are our best baseline for evaluation.

To reduce the shortcomings of an evaluation based on a strict string-based comparison (section 6), we propose a second type of evaluation: *semantic evaluation* (section 7). We then investigate in a third evaluation the potential value of automatically generated keywords. These can bring *new types* of search or archival behavior, that cannot be evaluated against current practices. For this we designed the in-use experiment serendipitous browsing (section 8).

But before presenting the evaluation methodologies and issues, let us introduce our automatic annotation pipeline (section 3), after a brief overview of such tools and platforms proposed in the literature (section 2).

2 Related Work

The tools and architectures that have been implemented for generating Semantic Annotations based on ontologies or other concept-based representation of a controlled vocabulary can be roughly categorized into:

- tools for manual annotation: an interface providing help for a human to insert semantic annotations in a text;
- tools for semi-automatic annotation: a system providing help and automatic suggestions for the human annotation;
- tools for automatic annotation: a system providing annotation suggestions, possibly to be validated or modified a posteriori.

Tools like Annotea (Kahan and Koivunen 2001) and SHOE (Heflin and Hendler 2000) provide environments for assigning *manually* annotations to documents; we aim at automatically suggesting them in our project, to ease some of the annotation burden.

The second category of tools proposes annotation suggestions after a learning process. They are represented by tools such as Amilcare (Ciravegna and Wilks 2003) and T-Rex(Iria 2005), that learn rules at annotation time in order to provide the annotator with suggestions. They are both based on the GATE platform (Cunningham et al. 2002), a generic Natural Language Processing platform that implements simple Named Entity recognition modules and a rule language to define specific patterns to expand on simple string recognition. Although we want to involve the Sound and Vision cataloguers in the annotation process, the cataloguers will ideally make use of our annotation suggestions to annotate AV programs, and not the context documents themselves. So, interactive annotation of the context documents is not the appropriate strategy to integrate semi-automatic annotation in the current process. Therefore, tools from the third category were considered the most relevant.

We opted for the Semantic Annotation performed by tools that generate them without human interaction. A typical example of this third type of tools is the KIM platform(Kiryakov A. and D. 2005); the MnM tool (Vargas-Vera M. and Fabio 2002) is mixed, providing both semi-automatic and automatic annotations. Although they can be adapted to different domains or use cases, the adaptation requires lots of work, and in the case of KIM, the upper level of the ontology cannot be changed. The annotation data model has to be integrated in the default structure provided by the tool, which would introduce tremendous changes in the thesaurus' structure, structure upon which we want to base our automatic annotations. The MnM system integrates an ontology editor with an Information Extraction pipeline, and this is also the approach that we decided to follow in our project, but we used GATE for this purpose, because of its openness and adaptability.

3 Annotation and Ranking Pipeline in the CHOICE-Project

Our approach to suggesting automatically keywords to annotate TV programs is based on Information Extraction techniques, applied to textual resources describing the TV program's content, such as TV-guide texts or web-site texts. Our system transforms these texts into a suggestion list of thesaurus keywords. The system comprises three parts:

1. A *text annotator*. The text annotator tags occurrences of thesaurus keywords in the texts. GATE(Cunningham et al. 2002) and its plug-in Apolda(Wartena et al. 2007) implement this process.
2. *TF.IDF computation*. For the TF.IDF we used $TF \times log(IDF)$. It ranks the keywords tagged in the previous stage.
3. A *Cluster-and-rank process* which uses the thesaurus relations to improve upon the TF.IDF ranked list.

See figure 1 for a schema of the total process.

The TF.IDF is a classic from Information Retrieval and is hard to improve upon. We use it as baseline which we try to beat with ranking algorithms. In this paper we only elaborate upon the cluster-and-rank algorithms. Our automatic process differs from the work performed by cataloguers. We only analyze associated text where cataloguers also inspect the original audiovisual material. Our automatic process generates a long list of **suggestions** where cataloguers **assign** a few keywords to a program.

3.1 Cluster-and-Rank Algorithms

The keywords tagged in the context documents of a TV program are sometimes related to each another by thesaurus relationships. Together the keywords

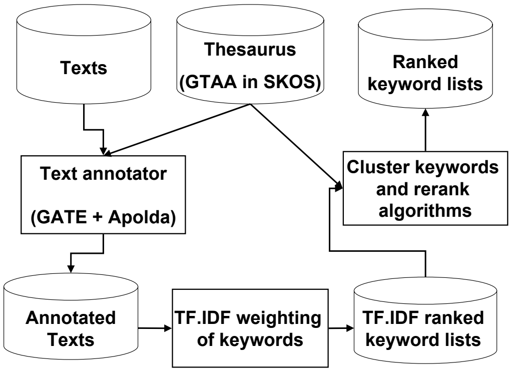
Texts
Thesaurus
(GTAA in SKOS)
Ranked keyword lists
Text annotator
(GATE + Apolda)
Cluster keywords and rerank algorithms
Annotated Texts
TF.IDF weighting of keywords
TF.IDF ranked keyword lists

Fig. 1. Schema of our system

and the relations form a graph. To increase the connectedness of our graph we also included indirect relations (in which an intermediate keyword connects two found keywords). The direct connections are defined as a relation of distance 1. Connections via intermediate terms are defined as relations of distance 2. An example is shown in figure 2.

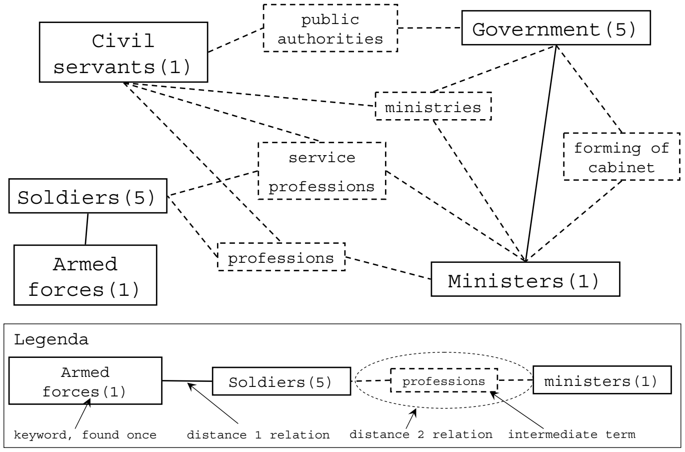

Fig. 2. Relations found between a set of keywords

The cluster-and-rank component uses the cluster structure of this connected graph to create a (re)ranked list as output. We implemented three algorithms that build ranked lists from this graph: the well known algorithm named Pagerank (Brin and Page 1998) (uses only graph information), our own method called CARROT (uses also TF.IDF information), and a second own method which is called Mixed (also uses TF.IDF information and the whole graph of the thesaurus as additional information).

CARROT CARROT (Malaisé, Gazendam, and Brugman 2007) stands for Cluster And Rank Related Ontology concepts or Thesaurus terms. It combines the local connectedness of a keyword and the TF.IDF score. The only graph property CARROT uses is the local connectedness of a keyword. It creates four groups each having the same local connectedness (group 1: both distance 1 and distance 2 connections, group 4: no connections). Each group is sorted on the TF.IDF values.

Pagerank Pagerank (Brin and Page 1998) is used to determine the centrality of items in a network and is used by Google. One way to understand the working of Pagerank is by imagining activation spreading through a network. The initial (*e.g.* TF.IDF) activation spreads itself equally via each available relations to other nodes in the network. It then spreads again via the relations of the network, some back to the original starting nodes and some further. In the end on each node in the network a dynamic equilibrium will be reached (each moment the same activation that leaves the node is also fed onto the node from other nodes, dynamic equilibrium). This equilibrium is no longer dependent on the starting activation, only on the network structure. The activation on each node corresponds with the Pagerank score and expresses its importance.

In research similar to our own by Wang et al. (Wang, Liu, and Wang 2007), Pagerank was used to determine the most central WordNet keywords in scientific articles. They compared Pagerank with TF.IDF and showed that Pagerank suggested much better keywords.

Pagerank is performed upon the same cluster as CARROT, but the Pagerank algorithm also assigns Pagerank scores to the intermediate terms so these are included in the suggestion list (also include the dashed terms of figure 2).

Mixed Algorithm using General Keyword Importance For the Mixed algorithm we wanted to keep some of the relevancy information conveyed by TF.IDF while performing the spreading of activation. We start with the TF.IDF activation and only spread it around with the official Pagerank formula during 3 iterations. At that moment some influence of the original TF.IDF is still present and at the same time some activation accumulates at the central nodes in the network. This Pagerank at t=3 is multiplied with the general importance of the keywords. The idea behind the weighting with keyword importance is that we want to favor keywords which are considered more important in general. The way we determine the general importance of keywords is by Pageranking the GTAA as a whole. We assume that the modeling of the GTAA reflects the importance of the keywords: topics which are considered important according to the GTAA makers from S&V are modeled with many keywords and many relations. The five keywords with the highest GTAA pagerank are *businesses, buildings, people, sports, animals*. The five keywords with the lowest GTAA pagerank are *lynchings, audiotapes, holography, autumn, spring*.

4 Source Material

Our corpus consist of 258 broadcasted TV-documentaries. 80% of these broadcasts belonged to three series of TV-programs: *Andere Tijden*, which is a series of Dutch historical documentaries, *Beeldenstorm*, which is a series of art documentaries presented by Henk van Os, the former director of the Rijksmuseum and *Dokwerk*, which is a series of historical political documentaries. Each broadcast is associated with one or more texts from the broadcasters web site (we name these *context documents*) and one manual catalogue descriptions made by S&V. The 258 TV-broadcasts are associated with 362 context documents. The length of the context documents varied between 25 words and 7000 words with an average of 1000 words.

4.1 Catalogue Descriptions

Each TV-broadcast in our corpus has a catalogue descriptions. These catalogue descriptions contain keywords which were assigned manually by cataloguers from S&V. The catalogue descriptions averagely contained 5,7 keywords with a standard deviation of 3,2 keywords. The minimum number of terms is 1, the maximum is 15. These keywords are the ground truth against which we evaluate the TF.IDF baseline and the three ranking algorithms in the next two experiments.

5 Experimental Setup

We perform three evaluations on two experiments. In our first experiment we generate keyword suggestions from contextual text for our corpus (see section 4) with the four different settings of our pipeline and we evaluate these against manually assigned keywords. We evaluate these resulting lists of suggestions in two different ways: classically and semantically.

Our first evaluation (section 6) is a classic Precision/Recall evaluation, inherited from the Information Extraction world. The task of suggesting keywords in the archival domain however made us question beforehand whether this classic evaluation methodology was appropriate given the reality of inter annotator disagreement.

The second evaluation (section 7) introduces a measure of semantic overlap between the Automatic Annotations and the target against we evaluate them: the manual annotations of the TV programs. This setting is still biased towards current annotation practices and do not show another dimension: what can Automatic Annotations bring in the context of possible *new applications*?

In order to evaluate the possibilities in terms of new practices in archives, we tuned a second experiment, which underlines the possible value of Automatic Annotations and Manual Annotations in the context of a particular search through an archive: Serendipitous Browsing (section 8). With it we test the value of the manual annotations and the CARROT keyword annotation suggestions for retrieving semantically related documents. By doing so, we feed an idea from the

Semantic Web (inherited from Semantic Browsing (Faaborg and Lagoze 2003) (Hildebrand 2008)) back into the archival world to bring new solutions to their core task: find relevant information/documents in large archives. Although the value of this idea needs to be tested, it reminds S&V's customer service of the loose search performed by users by flipping through a physical card-tray. The arrangement of physical cards in trays on one topic made it possible to browse for strong, semi or loosely related documents. This option was lost when the archives' access with card trays was replaced by computers.

6 Classical Evaluation

We want to measure the quality of the automatically derived keywords. For this purpose we compare the automatic annotations with the existing manual annotations. The standard way of evaluating our systems output against manual annotation is with the Information Retrieval measures of *precision* and *recall* (Salton and McGill 1983). **Precision** is defined as the number of relevant keywords suggested by our system for one TV-program divided by the total number of keywords that are given by our system for that program, and **recall** is defined as the number of relevant keywords suggested by our system for one TV-program divided by the total number of existing relevant keywords for that TV-program (which should have been suggested for that TV-program). Often precision and recall are inversely related, so it is possible to increase one at the cost of reducing the other. For this reason they are often combined into a single measure, such as the balanced F-measure, which is the weighted harmonic mean of precision and recall.

Given the fact that our system produces ranked lists, we can look at average precision and recall for different top parts our list: precision@5 and precision@10 express respectively the precision of the first 5 and the first 10 suggestion. For the suggestion of keywords to cataloguers only these top terms are important: a cataloguer will only read a limited number of suggestions. The cataloguer will stop when the suggestions are good (he has read enough good suggestions so he is satisficed (Simon 1957)) and stop when the suggestions are bad (he is not expecting reasonable suggestion anymore).

6.1 Classical Evaluation of the Results

Table 6.1 shows the classic evaluation for our four ranking algorithms.

The first observation we make is that only the Pagerank setting is considerably worse than the others. This is probably attributable to the fact that Pagerank lacks the ability to incorporate any relevancy information from the TF.IDF scores. The performance of Pagerank in the experiment of Wang(Wang, Liu, and Wang 2007) makes this result unexpected.

A second observation is that the Mixed model starts out as a very bad, but that it catches up with the better settings such as the TF.IDF baseline

precision		@1	@3	@5	@10
Baseline: TF.IDF	precision	0.38	0.30	0.23	0.16
CARROT	precision	0.39	0.28	0.22	0.15
Pagerank	precision	0.19	0.17	0.14	0.11
Mixed	precision	0.23	0.21	0.19	0.15
recall		@1	@3	@5	@10
Baseline: TF.IDF	recall	0.08	0.18	0.23	0.31
CARROT	recall	0.08	0.15	0.21	0.27
Pagerank	recall	0.04	0.09	0.13	0.20
Mixed	recall	0.05	0.12	0.18	0.28
F-score		@1	@3	@5	@10
Baseline: TF.IDF	F-score	0.13	0.22	0.23	0.21
CARROT	F-score	0.13	0.20	0.21	0.20
Pagerank	F-score	0.07	0.12	0.14	0.14
Mixed	F-score	0.08	0.16	0.19	0.20

Table 1. Classical Evaluation of our results

and CARROT. The TF.IDF seems best, but this difference is not statistically significant (at p ¡ 0.05).

A third observation is the big jump in F-score between @1 and @3 for all methods. This is interesting as it tells us that one suggestion just cannot contain that much information and that lists with 3 or 5 suggestions are better.

The final observation is that all the scores seem quite bad when we take in mind a representative performance by Dumais et al. (Dumais et al. 1998) with support vector machines on the well known Reuters-21578 dataset. The Reuters dataset is used for text classification and has 118 categories. Dumais et al. reach an optimal F-score on this set of 0.87. We however have more than 3800 different categories (keywords).

6.2 Discussion

Medelyan and Witten (Medelyan and Witten 2006) conducted an experiment similar to ours. They automatically derive keywords from the Agrovoc thesaurus (containing 16.600 preferred terms) for FAO documents (Food and Agriculture Organization of the United Nations). Their results show similar low numbers of around 0.20 for precision, recall and F-score. Their best method KEA++ reached the best F-score@5 of 0.187 with a precision@5 of 0.205 and a recall@5 of 0.197. Given that their documents are averagely 17 times longer than ours (which helps for retrieving good keywords) but that their number of possible keywords is 5 times as big too (which makes it harder to pick the right keyword), we can only state that our best methods produce reasonable results.

Inspection of individual suggestion lists reveals a mismatch between our sense of quality of the suggestions and the classic evaluation: many good suggestions do not contribute at all to the precision and recall numbers. To give an example: the first six CARROT suggestions for TV-program *Andere Tijden 04-09-2000* are *Jews, camps, deportations, interrogations, trains* and *boys*. The topic of this TV-program was the Dutch deportation camp of Westerbork from which Jews were deported to concentration camps in the second world war. The manual assigned keywords were *deportations, persecution of Jews, history* and *concentration camps*. According to the classic evaluation however only the suggestion of *deportations* is correct. Most of the other keywords however do convey valuable information. When we look at the relations of these suggested keywords in the GTAA, we see that *camps* is the broader term of *concentration camps* and that *Jews* is related to *persecution of Jews*. These thesaurus relations are used during semantic evaluation.

7 Semantic Evaluation

The classic type of evaluation takes place on the basis of exact match or *terminological consistency* (Iivonen 1995). We argue that this exact type of evaluation does not measure the quality of our suggestions well. We want keywords which present a semantic similarity with the manually assigned keywords to be counted as correct too. This is good enough for the task of suggesting keywords and it tackles part of the problem of the inter annotator disagreement. This semantic match is known as *conceptual consistency*(Iivonen 1995).

Medelyan and Witten (Medelyan and Witten 2006) describe a practical implementation of evaluation against conceptual consistency instead of terminological consistency. They use the relations in a thesaurus as a measure for conceptual consistency. The conceptually consistent terms are all terms which are within a certain number of thesaurus relationships from the target term. Medelyan and Witten consider in their experiment all terms reachable in two relations conceptually consistent (given their task and thesaurus). We chose to consider all terms within 1 thesaurus relationships to be conceptually consistent. This choice for 1 relationship is not purely motivated by the structure of our thesaurus, as it also would allow 2 steps of distance, but we face the risk of interaction between semantically based ranking methods (which use thesaurus relations) and the semantic evaluation methodology (which also uses thesaurus relations).

7.1 Results

We semantically evaluated the four settings against the manually assigned keywords. The results are presented in table 7.1.

In this table we see two things. First we observe from the F-scores that the Mixed setting is the best setting, but only @5 and @10. Its better F-score is only statistically significant @10. The Pagerank setting is again the worst setting, however it is only significantly worse than Mixed @5 and @10. The second observation is the difference in behavior with respect to precision and recall of the different methods. The Mixed model is good in precision, but normal in recall. CARROT is poor in recall and slightly better in precision.

When we compare table 6.1 and 7.1 we see a big improvement in performance. This not unexpected as the semantic evaluation effectively lowers the number of possible classes. We also see that the Mixed and the Pagerank setting improved much more than the other methods. Now we will look at the results qualitatively.

precision		@1	@3	@5	@10
Baseline: TF.IDF	precision	0.50	0.43	0.37	0.30
CARROT	precision	0.53	0.45	0.40	0.32
Pagerank	precision	0.47	0.40	0.36	0.30
Mixed	precision	0.52	0.46	0.42	0.36
recall		@1	@3	@5	@10
Baseline: TF.IDF	recall	0.16	0.32	0.40	0.54
CARROT	recall	0.17	0.28	0.36	0.48
Pagerank	recall	0.14	0.30	0.38	0.51
Mixed	recall	0.16	0.31	0.40	0.53
F-score		@1	@3	@5	@10
Baseline: TF.IDF	F-score	0.24	0.37	0.39	0.38
CARROT	F-score	0.25	0.35	0.38	0.39
Pagerank	F-score	0.22	0.34	0.37	0.38
Mixed	F-score	0.24	0.37	0.41	0.43

Table 2. Semantic Evaluation of our results

7.2 Qualitative Analysis

A qualitative analysis of the lists generated by the four different settings can give us some more insight into the value of the four ranking algorithms and into a possible interaction between semantic ranking methods and the semantic evaluation: does a setting scores well during the semantic evaluation because it is just a good setting, or because the evaluation prefers semantically connected keywords and the semantic settings (Pagerank, CARROT and Mixed) happen to suggest these. The TV-documentary *Andere Tijden: Mining accident at Marcinelle* is chosen for illustration.

Sound and Visions' catalogue describes this program as follows: *Episode of the weekly programme Andere Tijden. In this episode a mining accident in the fifties of last century in Belgium is addressed. In this mining accident many Italian foreign workers deceased during a fire.* The first 12 ranks generated by our four settings are displayed in table 7.2. The cataloguer attached the keywords *history, disasters, coalmines, miners* and *foreign employees* to this program. The catalogue keywords are not ranked (all are equally correct).

The keywords in **boldface** are exact matches with the catalogue keywords. The keywords in blue are conceptually consistent and the keywords in red are wrong.

From the table we make four observations. First we see that each list contains exactly three correct suggestions. In the TF.IDF and CARROT setting the keyword *miners, disasters* and *foreign employees* are in the list. The Pagerank and the Mixed setting have *miners* and *disasters* too, but they have *coalmines* as a third. Both the TF.IDF and the CARROT setting have many wrong suggestions in the list. The suggestion *mines* which is on top of the TF.IDF list, is wrong as it means an *under water bomb* in the GTAA. CARROT did not have this suggestion in the first group so it correctly is lower on the list. It also had *cables, safety* and *government* in a lower group.

Table 3. The suggested terms for Andere Tijden 2003-11-11: Mining disaster at Marcinelle

rank	TF. IDF	CARROT	Pagerank	Mixed	Catalogue
1	mines	**miners**	mines	mining	**history**
2	**miners**	**disasters**	mining	**miners**	**disasters**
3	**disasters**	fire	**coalmines**	**coalmines**	**coalmines**
4	fire	**forgn empl.**	publications	**disasters**	**miners**
5	cables	fathers	human body	accidents	**forgn empl.**
6	**forgn empl.**	corpses	buildings	blue-collar workers	
7	fathers	coal	art	coal	
8	corpses	mothers	**miners**	mines	
9	coal	firemen	accidents	fires	
10	safety	fires	families	families	
11	governments	immigrants	mining accidents	lignite	
12	mothers	immigration	**disasters**	golddiggers	

The Pagerank starts with three reasonable suggestions, but then from rank 4 until 7 gives very general suggestions. It favors suggestions that are very connected (and thus very general). The semantics of these suggestions is too general (not specific enough), which is often the case with the Pagerank suggestions. The following keywords appear in many of Pagerank's suggestion lists among the top ten: *publications, buildings, businesses, transportation, human body* and *professions*. If we would judge keywords within two relations as correct as Medelyan and Witten did, we would sometimes evaluate these general terms as correct.

The Mixed setting has a nice tradeoff between general suggestions and specific suggestions. It has some of the general suggestions like *mining* and *blue collar workers* which were introduced by Pagerank, but it also has suggestions specific enough to match the level of the usual manual annotations. Furthermore it has many more of the distance 1 suggestions in its list, not directly in the beginning, but further down the list. It does not generate more direct hits (table 6.1), but more semantic matches as table 7.1 shows. Mixed gives more closely related suggestions.

8 Serendipitous Browsing

After inspection of several lists of automatically derived keywords suggestions we discovered they contained four types. To illustrate the four types we again use the TV-program *Andere Tijden 04-09-2000* about the Dutch concentration camp Westerbork. The suggestion lists contain:

1. main topic descriptors *e.g. Jews, camps*
2. keywords related to the main topic *e.g. interrogations*
3. sub topic descriptors *e.g. trains*
4. wrong suggestions *e.g. boys*

The value of the first and non-value of the fourth type are clear. This second and third type would not be chosen by cataloguers to index a program, but they do convey interesting aspects of the program. Our lists of annotation suggestions contain exact suggestions, semantically related suggestions, sub topics and wrong suggestions. Lists belonging to two different broadcasts can contain the same keyword suggestion. This overlap can be used to link the broadcasts. Overlapping lists of annotation suggestions, although imprecise, might be a good measure of relatedness between two broadcasts. In the same manner overlapping manual annotations can relate two documents.

The value for users of these relations between documents can be great: to be able to browse through the archives, discover unsuspected relationships, thus creating new interpretations. It can create an accidental discovery or a moment of serendipity.

8.1 Experimental Setup for Serendipitous Browsing

We tested the value of the manual annotations and automatic annotations for serendipitous browsing with an experiment. For this experiment we created for our corpus a cross table, both for the manual annotations and for the automatic annotations in which we measure the overlap between documents. From both tables we selected the ten pairs with the biggest overlap. So we are cherry picking, but we did this with a reason. Our corpus contains only 258 programs, which represents only a small fraction of the entire catalog of over one million documents. For the entire catalogue we would get much better results. The best matches in our corpus give a better idea of what the method would mean for the entire catalogue.

For the automatic annotations the pairs had between 13 and 5 overlapping keywords. For the manual annotations these pairs had between 9 and 4 overlapping keywords. For each document in the top pairs we selected its four closest neighbors. This means that for each document A we have the five documents X1-X5 which have the highest number of overlapping keywords with document A. The first pair, A-X1 is one of the ten best pairs of either the manual annotations or the automatic annotations. The pair X1-A appears a second time as the first pair in the list of the five best pairs for document X1. The overlapping keywords for each pair represent the semantics of the link between the two documents.

In our list of results we identify three types of pairs:

1. The doc *X1* has a semantic overlap with doc *A*
2. X1 and A are two context documents of the same TV program
3. the documents *X1* and *A* constitute a part one and part two of a sequel

When pairs had a semantic overlap, we judged the similarity between the two documents on a five point Likert scale(Likert 1932): Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly Agree.

8.2 Results

The results are shown in table 8.2. Two documents appear twice in the list of 10 best manual annotation pairs. This means that a document is the most similar document for two different other programs. *Andere Tijden 2004-11-23 Rushdie affaire* is the most similar TV-program with respect to manual annotations for both *Andere Tijden 2003-09-30 Khomeiny* (with 5 overlapping keywords) and for *Andere Tijden 2005-02-01 The arrival of the mosque* (with 4 overlapping keywords).

Top 10	Automatic Annot	Manual Annot
linktype: documents have semantic overlap	4	5
linktype: error in database	1	1
linktype: two contextdocs form one TV-program	1	0
linktype: sequel part 1 and part 2	4	4
Unique documents in top 10 strongest links	20	18

Whole set	Automatic Annot	Manual Annot
Nb. of links	100	96
Nb. of semantic links	83	86
Nb. of unique semantic links	69	66
semantic link rating: Very good	5	2
semantic link rating: good	17	19
semantic link rating: neutral	31	27
semantic link rating: bad	8	26
semantic link rating: very bad	26	12
average link rating (1=very b, 5=very g)	2.59	2.66
average standard deviation in semantic rating	0.7	0.87
average nb. kw's	6.6	5.8
standard deviation Nb. kw's	2.3	2.1

Table 4. Typology of semantic links

It seems that the average quality of the semantic links is not very high: on average it tends slightly more to *neutral* than to *bad* for both sets. Given the small size of our corpus this is not very unexpected. It contains too few documents to generate many very good links. Still both the automatic annotations and the manual annotations have 21 *good* or *very good* semantic judgements. So with both annotations we could find quite some interesting links between documents. The do generate very different results however. Only eight of the pairs appear in both sets (8 out of 100), *i.e.* eight pairs were linked both via the manual annotations and the automatic annotations. Six of these constituted a part 1 and part 2 of a series. Both their catalogue descriptions and their context documents were much alike.

8.3 Qualitative Inspection

When we look at examples of semantic overlap we see very interesting results. We see for example that *Andere Tijden 2004-01-06* and *Andere Tijden 2004-12-07* get paired by the automatic annotations. The second program incorporated much of the content of the first program. According to the catalogue description the topic of the first program is: ”*the first Bilderberg-Conference which was held in 1954 under presidency of prince Bernhard*”. The topic of the second program is: "*the role prince Bernard played in the international circuit of politicians, soldiers and businessmen, especially his presidency of the international Bilderberg-meeting and his friendship with journalist Martin van Amerongen*". This second program was broadcasted just after the death of prince Bernard and incorporated much of the first programs material. The catalogue description does not mention the relation between the programs and the catalogue descriptions do not show a big overlap in terms of manual keywords. We manage to related these documents because the original makers adapted a context document of the first program and associated it to the second program. The automatic annotations derived from the original and the adapted context document show a big overlap. The manual annotations have only one overlapping keyword. The first program was indexed with the keywords *history, post-war rebuilding, secrecy, foreign policy, anti-Americanism, anti communism.* The second program was indexed with the keywords *history, conferences, politicians, entrepreneurs.* This difference is not only the result of the difference in the program. It serves as an example of inter annotator differences within the archives of Sound and Vision.

8.4 Discussion

Serendipitous browsing was created as a new way to evaluate the perceived value of the automatic annotations. We were not able to capture this value in the evaluation against manual annotations, neither in the exact evaluation nor in the semantic evaluation. However, the information specialists from S&V appreciated the new use of automatic techniques in a practical archive setting. In particular, the automatic linking of documents, whether it is done on the basis of manual annotations or automatic annotations, appears valuable and reminds of usages of the archive with the former physical card system. This linking of documents cannot be performed by hand (i.e., by human cataloguers) and lies outside the scope of the current archiving. An interesting result is the similar value for semantic browsing of automatic annotations compared to manual annotations: both sets of annotations generated the same amount of good and very good relations and on average both relations were judged with the same score. This suggests that although the automatic annotations are not as precise as the manual annotations, for semantic browsing purposes they have the same value.

9 Discussion and Perspectives

We set out to evaluate in three ways the value of automatic annotation suggestions for the audiovisual archive of S&V. The classic precision/recall evaluation showed that the baseline formed by TF.IDF ranking is the best ranking method. For the task of keyword suggestion within an archive however, this evaluation is too strict. The loosened semantic precision/recall measure showed that instead of the TF.IDF ranking the Mixed model performed best. As the Mixed model starts out worse then the TF.IDF, this result was only significant for the group of 10 first suggestions. The manual inspection showed that the Mixed tended to suggest more general terms. The third evaluation of manual and automatic annotations was in the Serendipitous Browsing experiment. It showed that the manual annotations and the automatic annotations have the same value for finding interesting related documents. With this experiment we only used the CARROT suggestions, so we are not able to differentiate ranking methods.

When we combine these three evaluation results and add to this the limited inter annotator agreement, it becomes hard to see how manual annotations can serve as *gold* standard. It is however the only material which we have. The question is how to evaluate against this resource and how to interpret the relevance of the outcome. As a first step it is good to apply semantic evaluation. A second step which we are working on is a user evaluation of our keyword suggestions by cataloguers from S&V. This user study is meant to have a human validation of the interest of the keywords suggestions for annotation and to get a deeper understanding of evaluation of our automatic keyword suggestion system.

As future work, we plan to experiment the suggestion of keywords based on automatic speech transcripts from the broadcasts and compare the results with the output generated from the context documents presented in this paper.

The interdisciplinary circle in this paper comes to a close: the practical archive setting forced us to change the classical way of evaluation and adapt novel ways of evaluation of our keyword suggestion system. The changed view on the evaluation however came back to the archive in the form of serendipitous browsing, which is perceived as a very interesting and probably valuable option for the daily archive. Even more interesting are our changed views: the problematic nature of evaluation changes the way we perceive Information Extraction and the archive has a radical new view on the future of archiving: it foresees that it will encompasses 80% automatic annotation and 20% manual annotation. Furthermore the thinking on automatic annotation will generate new ideas for interacting with the archive.

Our research follows a storyline often seen in the humanities, but uncommon for the sciences: instead of finding an improved solution to a known problem, as is common in the sciences, we got an almost Socratic understanding of evaluation: we now know that we have a very limited understanding of evaluation and only start to grasp the vastness of its problematic nature: we found problems and wonderment, as is common in the humanities.

References

Brin, Sergey, and Lawrence Page. 1998. “The anatomy of a large-scale hypertextual Web search engine.” *Computer Networks and ISDN Systems* 30 (1–7): 107–117.

Ciravegna, Fabio, and Yorick Wilks. 2003. Chapter Designing Adaptive Information Extraction for the Semantic Web in Amilcare of *Annotations for the Semantic Web*, edited by S. Handschuh and S. Staab, Volume 1, 112–127. IOS press.

Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. 2002. “GATE: A framework and graphical development environment for robust NLP tools and applications.” *Proceedings of the 40th Anniversary Meeting of the ACL.*

Dumais, Susan, John Platt, David Heckerman, and Mehran Sahami. 1998. “Inductive learning algorithms and representations for text categorization.” *CIKM ’98: Proceedings of the seventh international conference on Information and knowledge management.* New York, NY, USA, 148–155.

Faaborg, Alexander, and Carl Lagoze. 2003. “Semantic browsing.” *ECDL*, pp. 70–81.

Heflin, J., and J. Hendler. 2000. “Searching the web with shoe.” *Proceedings of the AAAI-2000 Workshop on AI for Web Search.* Budva, Montenegro.

Hildebrand, M. 2008. “Interactive Exploration Of Heterogeneous Cultural Heritage Collections.” *The Semantic Web - ISWC 2008*, Volume 5318 of *Lecture Notes in Computer Science.* 483 – 498.

Iivonen, M. 1995. “Consistency in the selection of search concepts and search terms.” *Information Processing and Management* 31 (2): 173–190 (March-April).

Iria, Jos. 2005. “T-Rex: A Flexible Relation Extraction Framework.” *proceedings of the 8th Annual CLUK Research Colloquium.* Manchester.

Kahan, J., and M.-R. Koivunen. 2001. “Annotea: an open RDF infrastructure for shared web annotations.” *World Wide Web.* 623–632.

Kiryakov A., Popov B., Terziev I. Manov D., and Ognyanoff D. 2005. “Semantic Annotation, Indexing, and Retrieval.” *Journal of Web Semantics* 2 (1): 49–79.

Leininger, K. 2000. “Inter-indexer consistency in PsycINFO.” *Journal of Librarianship and Information Science* 32 (1): 4–8.

Likert, Rensis. 1932. “A Technique for the Measurement of Attitudes.” *Archives of Psychology*, no. 140:155.

Malaisé, Véronique, Luit Gazendam, and Hennie Brugman. 2007. “Disambiguating automatic semantic annotation based on a thesaurus structure.” *14e conference sur le Traitement Automatique des Langues Naturelles (TALN).*

Medelyan, Olena, and Ian H. Witten. 2006. “Thesaurus-Based Index Term Extraction for Agricultural Documents.”

Salton, G., and M.J. McGill. 1983. *Introduction to modern information retrieval.* McGraw-Hill.

Simon, Herbert. 1957. *Models of Man.* Wiley New York.

Vargas-Vera M., Motta Enrico, Domingue John Lanzoni M. Stutt A., and Ciravegna Fabio. 2002. “MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup.” *In proceedings of the 13th Int.Conference on Knowledge Engineering and Management (EKAW-2002).* Siguenza, Spain.

Wang, Jinghua, Jianyi Liu, and Cong Wang. 2007. “Keyword Extraction Based on PageRank.” *Advances in Knowledge Discovery and Data Mining* 4426/2007:857–864.

Wartena, Christian, Rogier Brussee, Luit Gazendam, and Wolf Huijsen. 2007, September. "Apolda: A Practical Tool for Semantic Annotation." *The 4th International Workshop on Text-based Information Retrieval (TIR 2007).* Regensburg, Germany.