Search in audiovisual broadcast archives
Huurnink, B.

Citation for published version (APA):
Huurnink, B. (2010). Search in audiovisual broadcast archives

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Identifying Title and Thesaurus Terms in Queries

In Section 3.3.3, terms are identified by matching phrases and words within queries to metadata in clicked results. We include titles and thesaurus terms in the matching process, but exclude free text descriptions in the clicked results as these cannot be linked to any knowledge source. The resulting matched titles and entries are referred to as title terms and thesaurus terms respectively. The algorithm used to accomplish this is partially based on work by Snoek et al. [158] and outlined in Figure A.1. To summarize, a term is identified if: (1) at least one result is clicked during the search session in which the query is contained, (2) a candidate query phrase or word is contained in a program title or a thesaurus entry from a clicked result, and (3) there are not multiple candidate titles and/or entries conforming to the previous two conditions. As a result, not all terms can be identified, but coverage is extensive: of the 274,754 non-empty queries 203,685 were associated with at least one result click, and in turn 132,433 queries were associated with at least one title term or thesaurus term.

Manual Identification of Query Terms  To evaluate the quality and coverage of the automatically matched terms, an evaluation set was created. Here, three annotators manually identified the title terms and thesaurus terms contained in the queries from a randomly selected sample of 200 search sessions. The annotation procedure was set up as follows: the annotator was presented with the queries and clicked results from a search session, as well as a copy of the audiovisual thesaurus. The annotator was asked to identify the different title and thesaurus terms con-
Appendix A. Identifying Title and Thesaurus Terms in Queries

Input
Queries (all keyword searches)
Catalog entries with titles and thesaurus terms
Result click data mapping queries in a session to clicked catalog entries in the same session
Output
Set of query terms (titles and thesaurus entries from the clicked results that match or contain phrases within the queries)

Step 0: Preprocessing
- associate each thesaurus entry in the collection with any synonyms that may be contained in the thesaurus
- for each query, title, and thesaurus entry in the collection
  - strip punctuation, diacritics
  - remove frequently occurring stop words
  - stem words using the Porter stemming algorithm for Dutch (123)
  - split compound words using a compound splitter adapted from (109)

Step 1: Selection of Candidate Terms
- for each query
  - associate the query with clicked terms in the form of titles and thesaurus entries contained in the clicked session results
  - for each clicked term
    - count how many times the clicked term appears in the clicked results

Step 2: Processing of Query Phrases
- for each query
  - create a set of all possible phrases within the query that maintain the sequence ordering. (The longest phrase will be the entire query, the shortest phrases will be the individual words contained within the query)
  - order phrases by length, so that the phrase with the most words comes first
  - for each query phrase
    - initialize empty set of matched terms
    - if the query phrase is identical to (exactly matches) at least one clicked term
      - add all identical clicked terms to the set of matched terms
    - else, if the query phrase is contained in (phrase matches) at least one clicked term
      - add all container clicked terms to the set of matched terms
    - if the set of matched terms contains exactly one term
      - add the matched term to set of query terms
      - remove all query phrases overlapping the current phrase from processing
      - go to next query phrase
    - if the set of matched terms contains more than one term
      - select the matched terms that occur the most frequently in clicked results
      - if there is single matched term occurs most frequently in clicked results
        - add the single most matched term to set of query terms
        - remove all query phrases overlapping the current phrase from processing
        - go to next query phrase
      - if multiple terms occur most frequently in clicked results, the query term is ambiguous
        - remove all query phrases overlapping the current phrase from processing
        - go to next query phrase
    - go to next query phrase
  - go to next query

Figure A.1: Process to identify title terms and thesaurus terms contained in user queries. Phrases within a query are matched to candidate titles and thesaurus entries from clicked results. When a match is found, all words from the query phrase are removed from the term identification process.
Table A.1: Evaluation of automatic term matching using precision and recall, based on a sample of 356 queries from 200 sessions. # correctly matched indicates the number of automatically matched terms that were also manually identified, # matched indicates the total number of automatically matched terms, # correct indicates the total number of manually identified terms.

<table>
<thead>
<tr>
<th>Term source</th>
<th>Facet</th>
<th># correctly matched</th>
<th># matched</th>
<th># correct</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>n/a</td>
<td>108</td>
<td>114</td>
<td>157</td>
<td>0.95</td>
<td>0.69</td>
</tr>
<tr>
<td>Thesaurus</td>
<td>Genre</td>
<td>1</td>
<td>13</td>
<td>1</td>
<td>0.08</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>34</td>
<td>35</td>
<td>40</td>
<td>0.97</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Name</td>
<td>31</td>
<td>33</td>
<td>59</td>
<td>0.94</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Person</td>
<td>22</td>
<td>23</td>
<td>56</td>
<td>0.96</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>Program Maker</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Subject</td>
<td>42</td>
<td>42</td>
<td>86</td>
<td>1.00</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>All terms</td>
<td>n/a</td>
<td>241</td>
<td>267</td>
<td>406</td>
<td>0.90</td>
</tr>
</tbody>
</table>

tained in a query using the clicked results and the audiovisual thesaurus. When a term could not be identified in the thesaurus or the titles of the clicked results, the annotator was asked to mark this separately as an unknown term. Our sample of 200 search sessions contained a total of 356 queries; our annotators identified a total of 157 title terms, 249 thesaurus terms, and 109 unknown terms. The unknown terms were not contained in the audiovisual thesaurus or the titles, and included 21 dates and 38 codes that were entered by the user into the keyword search. The remaining unknown terms consisted of concepts not present in the thesaurus, such as night recording. We investigated the agreement between the three annotators on a hold-out set of 30 sessions using Krippendorf’s alpha [95]; we found the average pair-wise agreement to be 0.81.

Evaluation  The performance of our automatic term identification method on the manually annotated sample of sessions is shown in Table A.1. Performance is measured in terms of precision (the number of automatically identified terms that were labelled as correct, divided by the total number of automatically identified terms), and recall (the number of automatically identified terms that were labelled as correct, divided by the total number of terms that were labelled as correct). Overall, the automatically identified terms are accurate, with on average nine out of ten of the identified terms being correct. Some types of terms are easier to identify correctly than others; title terms and thesaurus terms from the Subject, Location, and Person facets all have a precision of over 0.95, while thesaurus terms from the Genre and Program Maker facets have a precision of less than 0.50. Recall is similarly variable, with over a relatively large proportion of the manually identified Location and Genre terms being returned by the automatic method. Less than one in two of the
manually identified terms are identified for the Person, Program Maker, and Subject facets.