Auomatic Annotation Suggestions for Audiovisual Archives

Luit Gazendam, Véronique Malaise, Annemieke de Jong, Christian Wartena, Hennie Brugman, Guus Schreiber

Content

- Context
- Project
- Related Work
- Annotation & Ranking
- Source Material
- Experimental Setup
- Evaluation
- Summary

Context

- Netherland Institute for Sound & Vision
  - archiving TV & radio programs (digitally)
  - Customer Groups:
    - Professional users from public broadcasters
    - Users from Science and Education
  - Typical Queries
    - Known query items
    - Subject queries
    - Shots and quotes
  - Annotation is bottleneck (manual)

Project

- Investigate how to automatically suggest keywords
- Evaluation:
  - String based
  - Semantic evaluation
  - Potential value
- Problem: Inter-cataloguer consistency 13% - 17% (ground truth?)

Related Work

1. tools for manual annotation
2. tools for semi-manual annotation
3. tools for automatic annotation

- tools from 3. most relevant for project

Annotation & Ranking Pipeline

1. text annotator
2. TF.IDF computation (ranking)
3. cluster-and-rank (improve TF.IDF)

Project	Cataloguers
only analyze associated texts	inspect original audiovisual material
generates list of suggestions	assign keywords

Cluster- and-Rank Algorithm

- Cluster terms with
  - Direct conntection (distance 1)
  - Intermediate term (distance 2)

- 3 algorithms:
  - Pagerank (uses graph info)
  - CARROT (uses TF.IDF info)
  - Mixed (uses TF.IDF info and whole graph of thesaurus)

Source Material

- 258 documentaries
- 362 context documents

for comparison:

- catalogue descriptions by cataloguers for each broadcast

Experimental Setup

1. generate keyword suggestions
   - classical and semantical evaluation
2. Serendipitous Browsing

Serendipitous Browsing

- Automatically derived keyword lists contain
  - main topic descriptures (good)
  - keywords related to main topic (value?)
  - sub topic descriptors (value?)
  - wrong suggestions (bad)

Serendipitous Browsing (ctnd.)

- cross table for manual / automatic annotations
- measure overlap of documents
- take ten pairs with greatest overlap
- identify
  - A and B have semantic overlap
  - A and B are context documents of same program
  - A and B constitute two parts of a sequel

Experiment 1 – Classical Eval.

- Precision: # of keywords suggested / # of keywords given
- Recall: # of keywords suggested / # of existing keywords

Usually: precision ~ 1 / Recall

-> F-Measure: weighted harmonic mean of precision and recall

Experiment 1 – Classical Eval.

- "Pagerank" considerably worse
- "Mixed" starts bad but catches up
- big jump from @1 to @3

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

Experiment 1 – Semantic Eval.

- conceptual consistency (≠ terminological consistency)

- "Mixed" is best for @5, @10
- "Pagerank" is worst
- "Mixed" good in precision, normal in recall
- "CARROT" poor in recall, better in precision

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

Experiment 2 - Evaluation

- automatic and mannual annotations have similar value

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

Summary

- for precision / recall classical “TF.IDF” is best
- for keyword suggestion “Mixed” is best
  - not as strict
- manual and automatic annotations have the same value for finding interesting related documents

Thank you for your attention!