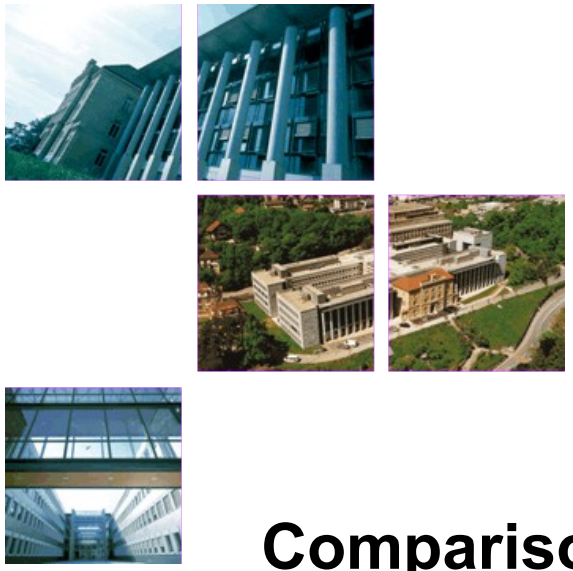

Université de Neuchâtel unine

Comparison Between Manually and Automatically Assigned Descriptors Based on a German Bibliographic Collection

Claire Fautsch, Jacques Savoy
Institute of Computer Science
University of Neuchâtel, Switzerland
{Claire.Fautsch, Jacques.Savoy}@unine.ch

Overview

- Introduction
- Test Collection
- IR Models and Indexing
- Manual Expansion
- Automatic Expansion
- Automatic Query Expansion
- Conclusion

Introduction

- Information retrieval (IR) from bibliographic records database
  - Title and abstract
  - Manually assigned Keywords
- Do manually assigned keywords improve retrieval?
- Do automatically assigned keywords eventually yield the same benefit?
- Does automatic query expansion help?

Example – Manually Added Keywords

- ACM Portal (abstract, reference, keywords)

↑ **INDEX TERMS**

Primary Classification:
H. Information Systems
↳ **H.3** INFORMATION STORAGE AND RETRIEVAL
↳ **H.3.3** Information Search and Retrieval
↳ **Subjects:** Clustering

Additional Classification:
H. Information Systems
↳ **H.3** INFORMATION STORAGE AND RETRIEVAL
↳ **H.3.3** Information Search and Retrieval
↳ **Subjects:** Retrieval models

General Terms:
Algorithms, Experimentation

Keywords:
Language modeling, aspect models, cluster hypothesis, cluster-based language models, clustering, interpolation model, smoothing

Example – Query Expansion

Test Collection

- German GIRT corpus
  - Social Science
  - 151,319 documents
- 125 queries
  - CLEF domain specific task
  - 3 parts (title, description, narration)
- Specific Thesaurus
  - German-English thesaurus for social science
- General Thesaurus
  - OpenThesaurus

```
<DOC>
<DOCNO>GIRT-DE19907042</DOCNO>
<DOCID>GIRT-DE19907042</DOCID>
<TITLE-DE>Vergleichende Studie zu Methoden der Auswertung von
kulturpolitischen Maßnahmen in Europa.
</TITLE-DE>
<AUTHOR>Bontinck, Irmgard</AUTHOR>
<AUTHOR>Angerer, Marie-Luise</AUTHOR>
<PUBLICATION-YEAR>1990</PUBLICATION-YEAR>
<LANGUAGE-CODE>DE</LANGUAGE-CODE>
<CONTROLLED-TERM-DE>Kulturpolitik</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>internationaler Vergleich</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Bewertung</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Forschungsbericht</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Erhebungsmethode</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Forschungsansatz</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Österreich</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Schweiz</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Jugoslawien</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Frankreich</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Schweden</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Europa</CONTROLLED-TERM-DE>
<METHOD-TERM-DE>empirisch</METHOD-TERM-DE>
<METHOD-TERM-DE>internationaler Vergleich</METHOD-TERM-DE>
<METHOD-TERM-DE>Aktenanalyse</METHOD-TERM-DE>
<METHOD-TERM-DE>Inhaltsanalyse</METHOD-TERM-DE>
<METHOD-TERM-DE>Sekundäranalyse</METHOD-TERM-DE>
<METHOD-TERM-DE>Querschnitt</METHOD-TERM-DE>
<CLASSIFICATION-TEXT-DE>spezielle Ressortpolitik
</CLASSIFICATION-TEXT-DE>
<ABSTRACT-DE>Vergleich von fünf Länderstudien zur jeweiligen
 Kulturpolitik. Inwieweit decken sich Anspruch und Realisierung,
welche Sparten werden als Kulturpolitik begriffen und behandelt?
Wie ist die methodische Vorgehensweise der Forschungsberichte?
Vorläufige Schlußfolgerung: der hohe theoretische Anspruch läßt
ich nur partiell einlösen. Die kulturellgesellschaftspolitische
Unterschiedlichkeit der Länder spiegelt sich in der
Art und Weise des untersuchten Feldes wider.</ABSTRACT-DE>
</DOC>
```

```
<top lang="de">
<num>10.2452/176-DS</num>
<title>Geschwisterbeziehungen</title>
<desc>Suchen Sie Dokumente, die die Entwicklung von Beziehungen
Zwischen Schwestern und Brüdern näher beschreiben.</desc>
<narr>Alle Dokumente, die die Beziehungen unter Geschwistern
in den verschiedenen Lebenslagen untersuchen sind relevant:
die Rolle von Geschwistern in der Familie, in der Schule, in der
Freizeit beziehungsweise die Veränderung der Beziehungen von der
Kindheit zum Erwachsenen sowie Unterschiede zwischen großen und
kleinen Familien.</narr>

</top>
<top lang="de">
<num>10.2452/177-DS</num>
<title>Arbeitslose Jugendliche ohne Berufsausbildung</title>
<desc>Suchen Sie Veröffentlichungen, die sich auf Jugendliche
beziehen, die arbeitslos sind und keine abgeschlossene
Berufsausbildung haben.</desc>
<narr>Relevante Dokumente geben einen Überblick über den
Umfang und die Probleme von Jugendlichen, die arbeitslos sind und
keine abgeschlossene Berufsausbildung haben. Nicht relevant sind
Dokumente, die sich ausschließlich mit Maßnahmen der Jugendhilfe
und Jugendpolitik beschäftigen.</narr>
</top>
```

```
<entry>
<german>Abbrecher</german>
<german-caps>ABBRECHER</german-caps>
<related-term>Abgänger</related-term>
<related-term>Aussteiger</related-term>
<related-term>drop out</related-term>
<english-translation>drop-out</english-translation>
</entry>
<entry>
<entry>
<german>Vergleich</german>
<german-caps>VERGLEICH</german-caps>
<scope-note-de>nicht im Sinne einer Regelung von
Rechtsstreitigkeiten, dann
Rechtsvergleich;</scope-note-de>
<narrower-term>internationaler Vergleich</narrower-term>
<narrower-term>interkultureller Vergleich</narrower-term>
<narrower-term>Modellvergleich</narrower-term>
<narrower-term>Ost-West-Vergleich</narrower-term>
<narrower-term>Soll-Ist-Vergleich</narrower-term>
<narrower-term>Leistungsvergleich</narrower-term>
<narrower-term>Kostenvergleich</narrower-term>
<narrower-term>Systemvergleich</narrower-term>
<narrower-term>Theorievergleich</narrower-term>
<narrower-term>regionaler Vergleich</narrower-term>
<narrower-term>Methodenvergleich</narrower-term>
<english-translation>comparison</english-translation>
</entry>
<entry>
<german>Verkehrsbelastung</german>
<german-caps>VERKEHRSBELASTUNG</german-caps>
<broader-term>Belastung</broader-term>
<broader-term>Umweltbelastung</broader-term>
<english-translation>traffic load</english-translation>
</entry>
```

IR Models and Indexing

- Indexing
  - Light stemmer
  - Decomposition
- IR Models
  - Vector space model (*tf idf*)
  - Language Model (LM)
  - Okapi
  - Divergence from Randomness (InB2)
- Evaluation
  - Mean Average Precision (MAP)

Manual Expansion

- Each document contains manually added keywords
- Include keywords into IR process
- Compare to baseline searching only in title and abstract

```
<CONTROLLED-TERM-DE>Kulturpolitik</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>internationaler Vergleich</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Bewertung</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Forschungsbericht</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Erhebungsmethode</CONTROLLED-TERM-DE>
```

Manual Expansion

	MAP		
Model	**Title & Abstract**	**+Manual**	**%Change**
tf idf	0.1929	0.2275	+17.94
LM	0.2865	0.3215	+12.22
InB2	**0.3157**	0.3493	+10.64
Okapi	0.3042	**0.3494**	+14.86
Mean	0.2673	0.3119	+13.91

Manual Expansion

- Keywords considerably improve overall retrieval performance
  - Improvement for 78 queries (InB2)
  - Decrease for 45 queries (InB2)
- Is it worth to spend human resources to add keywords?
  - Yes
  - But: Expensive in time and money
    - If not added by author
  - Can we achieve the same performance by automatically enhancing documents?

Automatic Expansion

- Keywords extracted from Thesaurus
  - Domain specific thesaurus
  - General thesaurus
- Select the *N* best matching terms
  - Jaccard similarity
  - $N = 50$

Automatic Expansion – Specific Thesaurus

	MAP		
Model	**Title & Abstract**	**+Manual**	**%Change**
tf idf	0.1929	0.1404	-27.22
LM	0.2865	0.1992	-30.47
InB2	**0.3157**	**0.2496**	-20.94
Okapi	0.3042	0.2151	-29.29
Mean	0.2673	0.2010	-24.80

Automatic Expansion – General Thesaurus

	MAP		
Model	**Title & Abstract**	**+Manual**	**%Change**
tf idf	0.1929	0.1874	-2.85
LM	0.2865	0.2380	-16.93
InB2	**0.3157**	0.2406	-23.79
Okapi	0.3042	**0.2654**	-12.75
Mean	0.2673	0.2328	-12.90

Automatic Expansion

- No average improvement
  - General thesaurus less impact than Specific
  - Less efficient then manual expansion
- Improvement on 22 queries compared to no expansion
  - Specific thesaurus and InB2 model
- Improvement on 22 queries compared to manual expansion
  - Specific thesaurus and InB2 model

Examples

- "Kinderlosigkeit in Deutschland"
  - No expansion: AP 0.5059
  - Manual expansion: AP 0.6030
  - Automatic expansion: AP 0.0839
- "Kinder- und Jugendhilfe in der russischen Föderation"
  - No expansion: AP 0.2894
  - Manual expansion: AP 0.0494
  - Automatic expansion: AP 0.1930

Automatic Query Expansion

- User not knowledgeable in the domain
  - Does not use specific vocabulary
  - Manual expansion not possible
  - Automatic expansion using thesaurus
- Jaccard similarity
- 5 best matching terms
- Short (title only) and long (title and description) queries

Automatic Query Expansion
Short Queries

	MAP			MAP	
Model	**No Exp.**	**GIRT**	**%Change**	**OpenThes**	**%Change**
tf idf	0.2275	0.2285	+0.44%	0.2289	+0.62%
LM	0.3215	0.3240	+0.78%	0.3233	+0.56%
InB2	0.3493	0.3485	-0.23%	0.3483	-0.29%
Okapi	**0.3494**	**0.3503**	+0.26%	**0.3510**	+0.46%
Mean	0.3119	0.3128	+0.28%	0.3128	+0.28%

Automatic Query Expansion
Long Queries

	MAP			MAP	
Model	**No Exp.**	**GIRT**	**%Change**	**OpenThes**	**%Change**
tf idf	0.2428	0.2430	0.08	0.2431	0.12
LM	0.3606	0.3621	0.42	0.3616	0.28
InB2	0.379	0.3795	0.13	0.3793	0.08
Okapi	0.3856	0.3861	0.13	0.3865	0.23
Mean	0.3420	0.3427	0.20	0.3426	0.17

Automatic Query Expansion

- Only small mean improvements
  - Mainly due to change of order of items retrieved
- Improvement for 52 queries, decrease for 72
  - Short queries
  - InB2
  - Specific Thesaurus
- Improvement for 36 queries, decrease for 38
  - Short queries
  - InB2
  - General Thesaurus

Examples

- “Radio und Internet”
  - No expansion: AP 0.3986
  - Specific thesaurus: AP 0.4509
    - Rundfunk, Datennetz, Datenaustausch, Welle
  - General thesaurus: AP 0.4846
    - Hörfunk, Rundfunk, Netze, Netz, Funk

Examples

- “Generationsunterschiede im Internet”
  - No expansion: AP 0.4789
  - Specific thesaurus: AP 0.4408
    - Datennetz, Datenaustausch, Intranet
  - General thesaurus: AP 0.4827
    - Netze, Netz, Web, WWW, World

Conclusions

- Manually added keywords improve retrieval effectiveness
- Worth spending human resources
  - Take into account context
  - Chose appropriate Keywords
- Automatic document expansion fails
  - Context

Conclusions

- Query expansion shows only small improvements
  - Important variations depending on the query
  - Unforeseeable when it helps and when it doesn't

Questions???