How One Word Can Make all the Difference

Using Subject Metadata for Automatic Query Expansion and Reformulation

Vivien Petras
School of Information Management & Systems
UC Berkeley

Overview

- Introduction to idea
- Results
- Examples for advantages and problems
- Thesaurus expansion vs. blind feedback
- Further work
- Bilingual

Subject Metadata - Purpose

Subject-describing keywords: thesauri, classification systems, subject heading list, ontologies

= **controlled vocabularies**

- Concise topical description of content
- Non-ambiguous term for each concept represented
- All relevant docs for a concept under one term
- More searchable text for the searcher

Subject Metadata for Retrieval

- Premise: Using subject metadata can lead to generally shorter, more precise and more complete searches
- Controlled vocabulary terms may differ from searcher vocabulary

  →Additional learn / search effort
- Approach: automatic suggestion of controlled vocabulary terms for query expansion & reformulation

The GIRT Collection

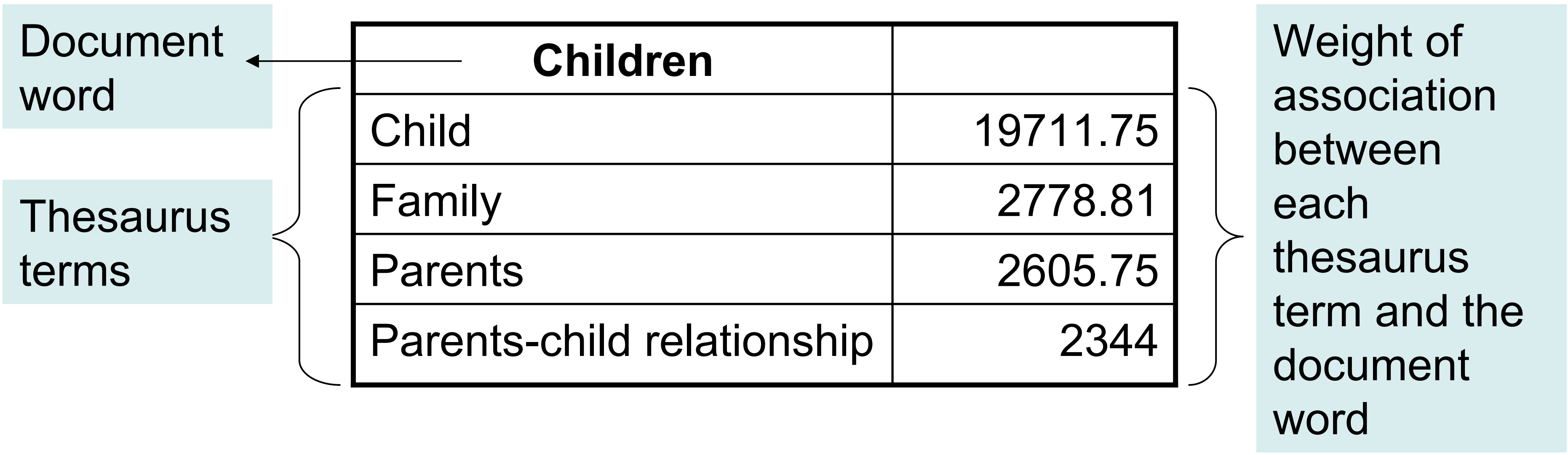
- 150,000 documents in 2 collections (English & German) in the social science domain
- Titles, abstracts and 10 thesaurus terms (phrases) per document
- Ca. 7,000 unique thesaurus terms / phrases

Thesaurus terms are not evenly distributed:

- Most occur <100 times ➔Good for retrieval
- 307 terms occur >1,000 times
- “Bundesrepublik Deutschand” occurs 60,955 times

Entry Vocabulary Modules for GIRT

- Words and thesaurus terms that are related will co-occur more often
- Co-occurrence matrix between title/abstract words and thesaurus terms in documents
- Each word/descriptor pair is assigned an association rank (weight)



Document word

Thesaurus terms

Children	
Child	19711.75
Family	2778.81
Parents	2605.75
Parents-child relationship	2344

Weight of association between each thesaurus term and the document word

Query Expansion

- Query expansion: look up each query title word in EVM, add 2 highest ranked suggested thesaurus terms to the query

	138	143
Title	**Insolvent Companies**	**Giving up Smoking**
EVM	*Liquidity / Indebtedness* *Enterprise / Firm*	*Donation / Social relations* *Smoking / Tobacco consumption*

Monolingual Retrieval

- Comparing baseline run (TD) with best query expansion run:

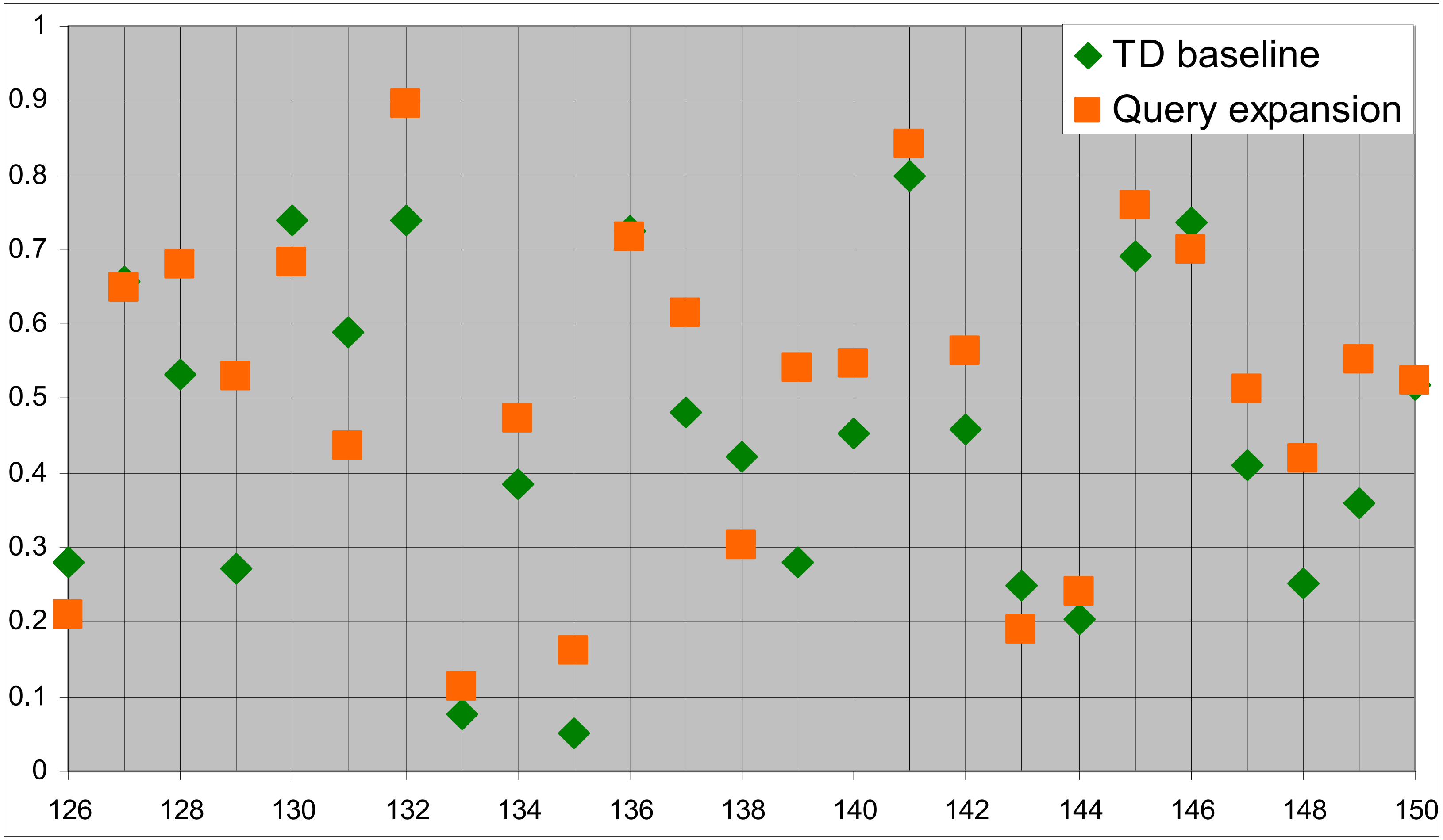
	TD baseline	TD with query expansion
German	0.4547	0.5144 (+13%)
English	0.4531	0.4818 (+6%)

- Comparing on a query-by-query basis:

	TD baseline	TD with query expansion
German	8	17
English	9	16

- Query expansion improves retrieval in 2/3 of the cases.

German Monolingual Retrieval



Most improvement: 135 (+210%), 129 (+ 94%), 139 (+92%)

Most decrease: 138 (- 28%), 131 (-26%)

The Impact of Individual Words

Example query 139:

Title: Gesundheitsökonomie (Health Economics)

Desc: Finde Dokumente, die die Versorgung der Bevölkerung mit medizinischen und ärztlichen Dienstleistungen aus ökonomischer Sicht diskutieren.

EVM suggestions: Gesundheitswesen / Ökonomie (Health care delivery system / Economy)

TD baseline	0.2812
TD + Gesundheitswesen	0.3751 (+ 33%)
TD + Ökonomie	0.4056 (+44%)
TD + Gesundheitswesen + Ökonomie	0.5049 (+79%)

The Impact of Individual Words

Example query 131:

Title: Zweisprachige Erziehung
(Bilingual Education)

Desc: Finde Dokumente, die die bilinguale Erziehung diskutieren.

EVM suggestions:
Mehrsprachigkeit / interkulturelle Erziehung
(Multilingualism/ intercultural education)
Erziehung / Pädagogik
(Education / Pedagogics)

TD baseline	0.5901
TD + EVM suggestions	0.4371 (-26%)
TD + EVM - Erziehung	0.5676

The term „Erziehung“ is too common to add valuable information to the query and also already occurs in the query.

The Impact of Individual Words

Example query 129:

Title: Sexualität und Behinderung
(Sexuality and Disability)

Desc: Finde Dokumente, die das Thema Sexualität und Behinderung diskutieren.

EVM suggestions:
Sexualität / Homosexualität
(Sexuality / Homosexuality)
Behinderung / Behinderter
(Handicap / Handicapped)

TD baseline	0.2729
TD + Sexualität	0.2792
TD + Homosexualität	0.2925
TD + Behinderung	0.4018 (+47%)
TD + Behinderter	0.4692 (+72%)
TD + all EVM	0.5295

Just removing thesaurus terms that already occur in the query might not be the best strategy.

EVM vs. Blind Feedback

- Pre-retrieval
- Adds terms from controlled vocabulary
- Adds 2-6 terms

- Post-retrieval
- Adds terms from highest ranked documents
- Adds 20 terms from top 30 documents

	TD baseline	Blind feedback	EVM
German	0.4622	0.4547 (9)	0.4902 (16)
English	0.4175	0.4531 (16)	0.4517 (13)

Short Queries – Title only

- Very effective for short queries: adding EVM terms to title-only queries improves precision more than blind feedback
- Thesaurus terms are downweighted in order not to dominate the query words

	T baseline	Blind feedback	EVM	Blind feedback + EVM
German	0.4030	0.3643 (9)	0.4522 (17)	0.4748 (22)
English	0.3415	0.3972 (18)	0.4140 (19)	0.4542 (18)

Summary / Further Work

- Query expansion with controlled vocabulary words improves retrieval
- For English: better phrase processing (German compounds don‘t split concepts)
  - Gesundheitswesen ↔ Health care delivery system
  - For EVM matching and retrieval
- Individual words have a big impact
  - High-quality search terms (not too broad or vague)
  - Query analysis: only add new / non-common words?

Bilingual Retrieval

- Machine translation of TD
  - combined Systran and L&H Power Translator
- Multilingual Thesaurus:
  - query title words were submitted to EVM in source language, suggested terms were replaced with thesaurus terms in target language

	Machine Translation	EVM Thes. Terms	Combined
German → English	0.3917 (14)	0.3339 (11)	0.4566 +17%
English → German	0.3532 (15)	0.3236 (10)	0.4059 +15%

- Thesaurus-terms-only performs almost as well as machine translation
- Combining the techniques almost achieves monolingual performance