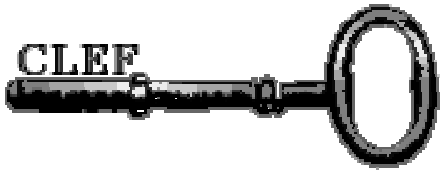
Bibliographic database access using free text and controlled vocabulary



Jacques Savoy

University of Neuchatel, Switzerland

www.unine.ch/info/clef/

Bibliographic database

- Controlled vocabulary vs. free text
- Combined strategy
- Thesaurus

Amaryllis collection

<DOCID> AM-000002
<TEXT>
<TI> Dermatite atopique : Actualisation
<AB> La dermatite atopique (DA) est une pathologie fréquente du jeune enfant, dont le traitement classique repose sur des soins locaux adaptés. ...

<MC> Dermatite atopique, Enfant, Nourrisson, Article synthèse
<KW> Atopic dermatitis, Child, Infant, Review

Amaryllis collection

148,688 articles (195 MB)

with <ab>, <mc> and <kw>

“only” 110,528 (74%) with also <ti>

25 queries with in mean

80.72 relevant documents per request

Stemming in French

Removing inflectional suffixes

masc sing	l'ami
fem sing	l'amie
masc plur	les amis
fem plur	les amies

and some derivational suffixes

CLEF-01 vs. CLEF-02

CLEF-02 collection, Query=TD, Okapi

Setting	CLEF-01	CLEF-02	
French	43.51	47.12	(+8.3%)
(stoplist & stemmer)			
“optimum”		48.41	(+11.3%)
(Okapi parameters)			

Retrieval models

binary {0, 1} (bnn)

tf = occurrence frequency (nnn)

idf = inverse document frequency

tf · idf, (ntc)

(log(tf)+1) · idf, (ltn)

log(log(tf)+1)+1) · idf, (dtc)

(Singhal et al., TREC-7)

Okapi (Robertson et al., IPM, 2000)

10 different retrieval models

Controlled vocabulary evaluation

<kw>&<mc>	Title	TD	TDN
bnn-bnn	22.7	19.8	24.6
nnn-nnn (tf)	8.6	11.0	13.5
ntc-ntc (tf·idf)	17.6	24.2	28.3
ltn-ntc	26.4	32.9	39.3
dtu-dtn	28.5	32.3	40.4
Okapi	**29.8**	**38.1**	**45.4**

Controlled vocabulary evaluation

Longer queries obtain better results

	in mean	
Title	TD	TDN
baseline	+23%	+53%

Controlled vocabulary vs. free text

	Title <kw&mc>	Title <ti&ab>	
bnn-bnn	22.7	11.3	(-50.3%)
nnn-nnn (tf)	8.6	5.1	(-40.7%)
ntc-ntc (tf·idf)	17.6	16.0	(-8.8%)
ltn-ntc	26.4	20.4	(-22.7%)
dtu-dtn	28.5	23.9	(-16.2%)
Okapi	**29.8**	**23.9**	(-19.6%)

Controlled vocabulary vs. free text

Controlled vocabulary seems to provide a better performance than free text indexing (difference around 20%).

Using T, TD or TDN, the conclusion is the same.

Controlled vocabulary and free text indexing

	Title <kw&mc>	Title <all>	
bnn-bnn	22.7	21.0	(-7.4%)
nnn-nnn (tf)	8.6	9.0	(+4.2%)
ntc-ntc (tf·idf)	17.6	21.6	(+22.6%)
ltn-ntc	26.4	31.8	(+20.4%)
dtu-dtn	28.5	31.8	(+11.6%)
Okapi	**29.8**	**36.3**	(+22.0%)

Controlled vocabulary and free text indexing

Combining controlled vocabulary and free text indexing improves the retrieval effectiveness (around 18%, T, TD or TDN)

- Blind query expansion may improve this performance (+7%)

User's point of view

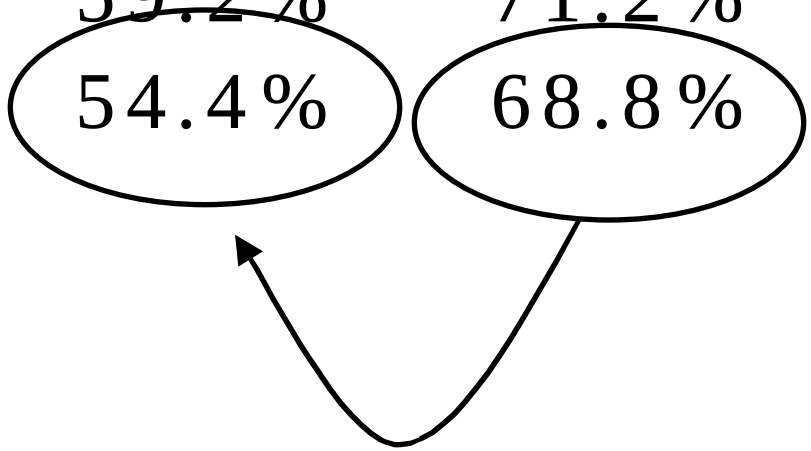
	Okapi model		
	<kw&mc>	<ti&ab>	<all>
Prec@5	60.8%	59.2%	71.2%
Prec@10	54.0%	54.4%	68.8%

And if we give up manually indexing?

User's point of view

Okapi model

	<kw&mc>	<ti&ab>	<all>
Prec@5	60.8%	59.2%	71.2%
Prec@10	54.0%	54.4%	68.8%



Thesaurus

173,946 entries <TERMFR>
173,946 entries <TRADENG>
26,160 entries <SYNOFRE>
28,801 entries <AUTOP1>
1,937 entries <VAUSSI1>

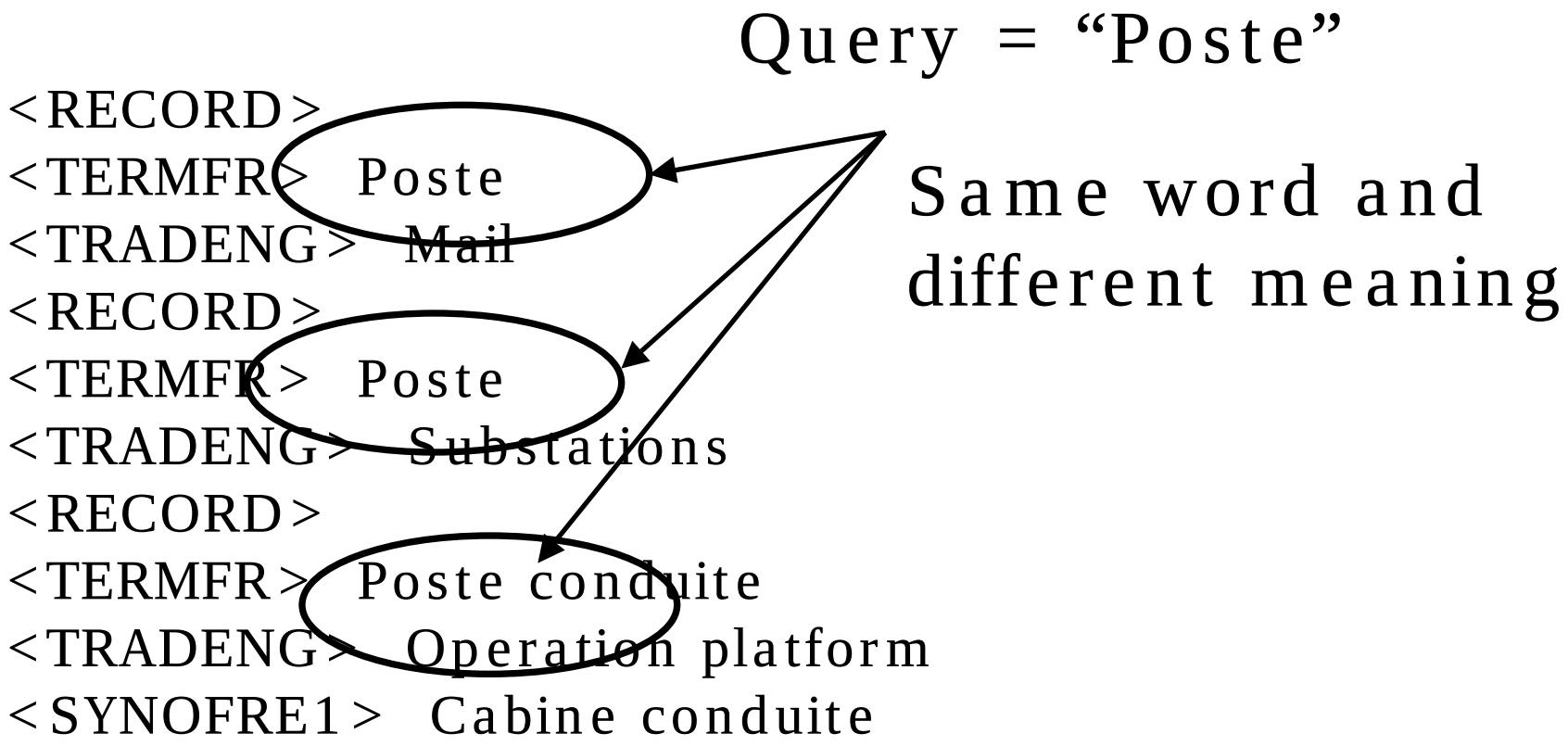
<RECORD>
<TERMFR> Poste
<TRADENG> Mail
<RECORD>
<TERMFR> Poste
<TRADENG> Substations
<RECORD>
<TERMFR> Poste conduite
<TRADENG> Operation platform
<SYNOFRE1> Cabine conduite

Thesaurus

Number of entries composed
1 word: 35,709 (28.1%)
2 words: 65,594 (51.7%)
3 words: 20,438 (16.1%)

<RECORD>
<TERMFR> Poste
<TRADENG> Mail
<RECORD>
<TERMFR> Poste
<TRADENG> Substations
<RECORD>
<TERMFR> Poste conduite
<TRADENG> Operation platform
<SYNOFRE1> Cabine conduite

Thesaurus expansion



Conclusion

- Okapi is a good IR model
- Controlled vocabulary seems to be better than free text,
- However, combining is the best way
- Thesaurus expansion?