



The rise and fall of stopword lists in patent searching and mapping



PIUG Annual Conference – 05/2006
Antoine Blanchard – Patent Information Team, Syngenta

Historical background (1/3)

Hans Peter Luhn (1896-1964)

- pioneer in indexing and information retrieval
- International Conference on Scientific Information, Washington DC (1958): introduced the **Keyword-in-Context** indexing technique (KWIC)
- first stopwords

```
truth universally acknowledged, that a single  
possession of a good fortune, must be in want of  
possession of a good fortune, must be in want  
of, that a single man in possession of a good  
fortune, that a single man in possession of a good  
fortune, must be in want of a wife.  
It is a truth universally acknowledged,  
that a single man in possession of a good  
fortune, must be in want of a wife.  
It is a truth universally acknowledged,  
that a single man in possession of a good  
fortune, must be in want of a wife.
```

Historical background (2/3)

C.J. van Rijsbergen (1943-)

- 1975: *Information Retrieval* suggests 250 stopwords
→ form the now classical default stopword list

a	after	all	that	themselves
about	afterwards	almost	the	then
above	again	(...)	their	thence
across	against	than	them	(...)

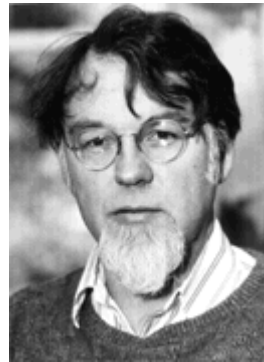
Historical background (3/3)

Since 1975: no improvement!

1958: KWIC,
first stopwords

1975: classical van Rijsbergen's
stopword list

2006



Luhn, 1952 (courtesy IBM)

Stopword lists in patent information today (1/3)

In databases:

- stopwords used *a minima*
 - e.g. in CA: “an”, “and”, “as”, “at”, “by”, “for”, “from”, “in”, “not”, “of”, “on”, “or”, “the”, “to”, “with”
 - *do* are used in full text databases (318 stopwords in Micropatent, 90 in USPATFULL...)
- since 2000's, many database reloaded → obliteration
 - “stopwords are no longer used” in NTIS, FSTA, WPIX...
- no longer need for stopwords in retrieval?!

Stopword lists in patent information today (2/3)

In patent mapping and clustering tools:

- purpose:
 - algorithms based on the lexical content of the documents to map them accordingly
 - remove stopwords → keep only information-bearing words
- content:
 - STN[®] AnaVist[™]: proprietary stopwords list
 - Thomson Aureka[®]: 1,290 stopwords ~ comprising van Rijsbergen's + words in German and French
 - ↳ doesn't even completely cover the 'patent language'!

Stopword lists in patent information today (3/3)

Rise and fall of stopwords lists:

- stopwords deprecated and obliterated, despite the revival in mapping
- this trend can be seen as a progress for the searcher... but not for stopwords themselves!

However, room for improvement in patent mapping

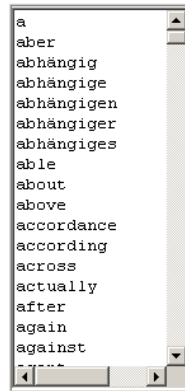
→ toward reestablishment of stopwords lists in patent mapping?

Customizing stopwords lists 101

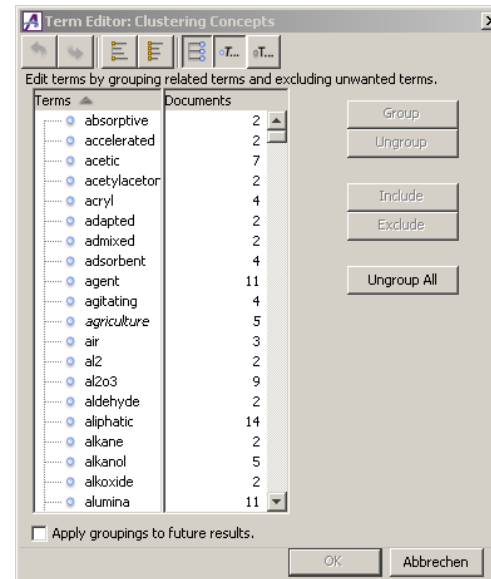
Stopwords:

Stopwords are words excluded during map creation. To add, type words in any order, separated by spaces or on separate lines.

Note: To store custom stopwords sets, copy and paste into/from text files.



a
aber
abhängig
abhängige
abhängigen
abhängiger
abhängiges
able
about
above
accordance
according
across
actually
after
again
against



Add 'patent language' terms:

- “embodiment”, “comprising”, “providing”, “relating”, “device”, “capable”, “desired”, “preferred”, “example”, “exhibiting”...

Add more?

- use this feature to enhance mapping...

Customizing stopwords lists 102 (1/2)

Add words that are information-bearing but trivial in a given context

- for example, in a corpus of documents dealing with the use of phytase enzyme as feed additive, the word “protein” becomes trivial

How?

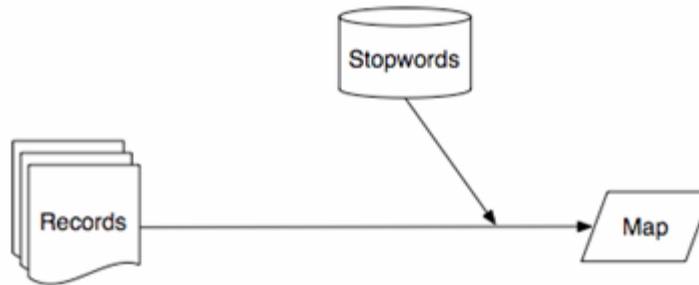
- trial and error: first round, second round,... stop when satisfied

When?

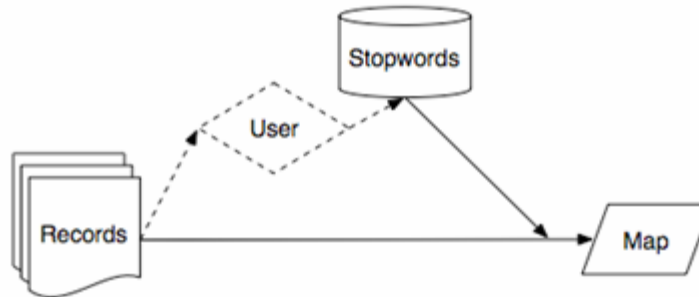
- for discriminating between clusters
- for skewing the analysis in the direction desired

Automated stopwords list construction (1/3)

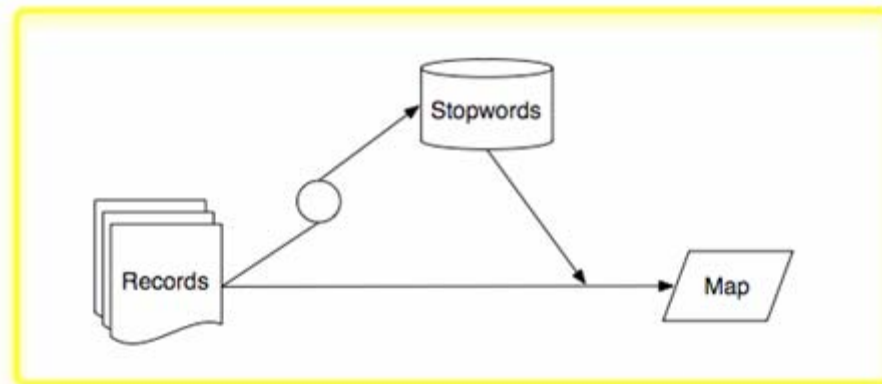
Preset stopwords list



User-defined stopwords list



Automated stopwords list construction



Automated stopwords list construction (2/3)

Featured in AnaVist

- for example, in the record 2004:1057892 related to the xenograft of ovarian tissue into nude mice, the terms “collection”, “development”, “influence”, “young”... are considered as stopwords
- customizable parameter called “Concept Frequency”
- one can add but not remove stopwords
- proprietary algorithm → ‘black box’

Featured in OmniViz

- advanced lexical handling
- stopwords + so-called “Other Terms”
- supports multi-word phrases, synonyms...
- proprietary algorithm but fully editable lists

Automated stopwords list construction (3/3)

Other solutions, to be implemented by analysts or (pref.) vendors

Examples of algorithms:

- term-based random sampling, optionally merged with a classical stopwords list (Lo *et al.*, *Automatically building a stopwords list for an information retrieval system*, 2005 http://ir.dcs.gla.ac.uk/terrier/publications/rtlo_DIRpaper.pdf)
- evolutionary algorithms (Sinka & Corne, *Evolving Better Stoplists for Document Clustering and Web Intelligence*, 2003 <http://www.sse.reading.ac.uk/common/publications/02065.pdf>)

Acknowledgements...

Tony Trippe & Brian Sweet, CAS

Jeff Saffer, OmniViz