Introducing the User-over-Ranking hypothesis

Benno Stein Matthias Hagen

Bauhaus-Universität Weimar
matthias.hagen@uni-weimar.de

ECIR 2011
Dublin, Ireland
April 21, 2011

User-over-Ranking stated informally

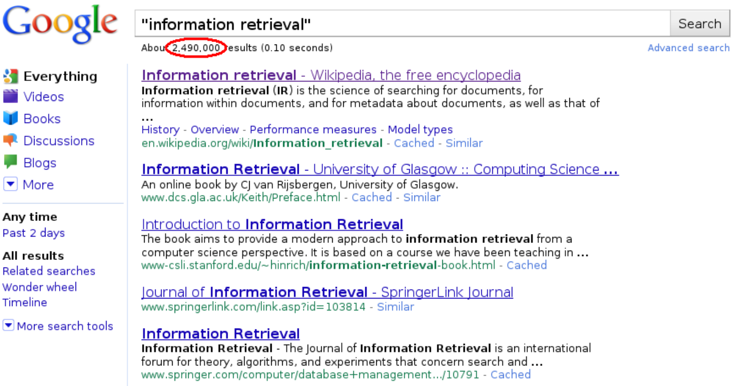
Queries returning as many results as the user can consider increase retrieval performance.

User-over-Ranking stated informally

Queries returning as many results as the user can consider increase retrieval performance.

Small print: If ranking works: fine!
Use case is not some query like ebay.
But more involved information needs,
automatic systems, etc.

Assumption 1: More keywords = more specific



"information retrieval"

About 2,490,000 results (0.10 seconds)

Information retrieval - Wikipedia, the free encyclopedia
Information retrieval (**IR**) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of ...
History - Overview - Performance measures - Model types
en.wikipedia.org/wiki/**Information_retrieval** - Cached - Similar

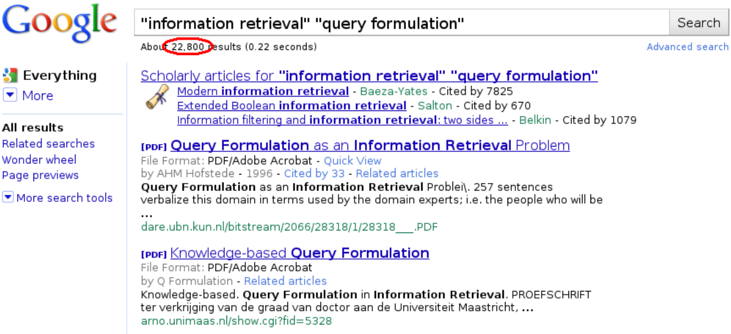
Information Retrieval - University of Glasgow :: Computing Science ...
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - Cached - Similar

Introduction to **Information Retrieval**
The book aims to provide a modern approach to **information retrieval** from a computer science perspective. It is based on a course we have been teaching in ...
www-csli.stanford.edu/~hinrich/**information-retrieval**-book.html - Cached

Journal of **Information Retrieval** - SpringerLink Journal
www.springerlink.com/link.asp?id=103814 - Similar

Information Retrieval
Information Retrieval - The Journal of **Information Retrieval** is an international forum for theory, algorithms, and experiments that concern search and ...
www.springer.com/computer/database+management.../10791 - Cached

Assumption 1: More keywords = more specific



"information retrieval" "query formulation"

Search

About 22,800 results (0.22 seconds)

Advanced search

Everything
More

All results
Related searches
Wonder wheel
Page previews

More search tools

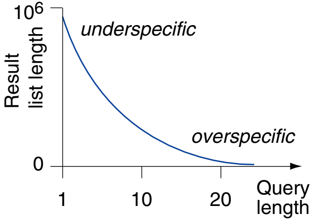
Scholarly articles for "information retrieval" "query formulation"

Modern information retrieval - Baeza-Yates - Cited by 7825
Extended Boolean information retrieval - Salton - Cited by 670
Information filtering and information retrieval: two sides ... - Belkin - Cited by 1079

[PDF] Query Formulation as an Information Retrieval Problem
File Format: PDF/Adobe Acrobat - Quick View
by AHM Hofstede - 1996 - Cited by 33 - Related articles
Query Formulation as an **Information Retrieval** Problei\. 257 sentences verbalize this domain in terms used by the domain experts; i.e. the people who will be ...
dare.ubn.kun.nl/bitstream/2066/28318/1/28318___.PDF

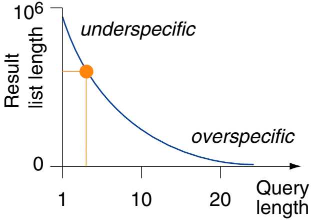
[PDF] Knowledge-based Query Formulation
File Format: PDF/Adobe Acrobat
by Q Formulation - Related articles
Knowledge-based. **Query Formulation** in **Information Retrieval**. PROEFSCHRIFT ter verkrijging van de graad van doctor aan de Universiteit Maastricht, **...**
arno.unimaas.nl/show.cgi?fid=5328

Assumption 1: More keywords = more specific



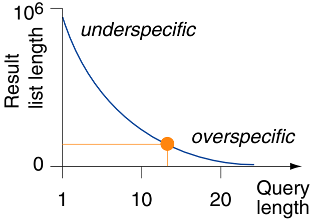
Specificity of Queries

Assumption 2: User can arbitrarily specify information need



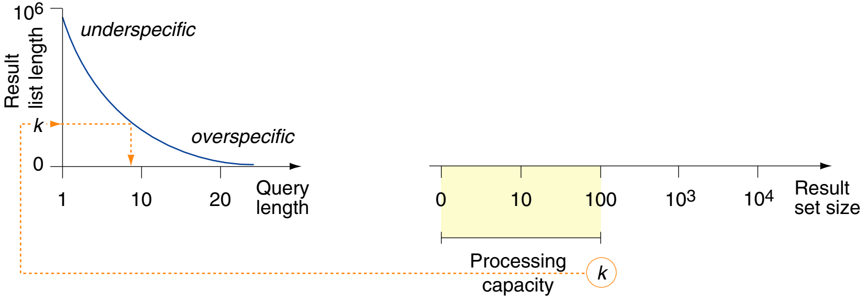
Specificity of Queries

Assumption 2: User can arbitrarily specify information need



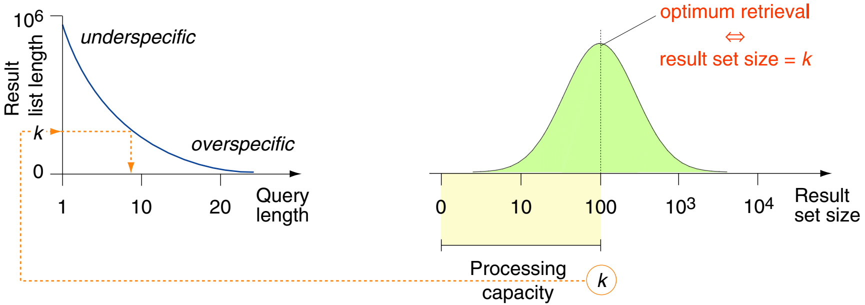
Specificity of Queries

Assumption 3: User can consider about k results.



Specificity of Queries

Hypothesis: Specificity matches k = Optimum retrieval



Specificity of Queries

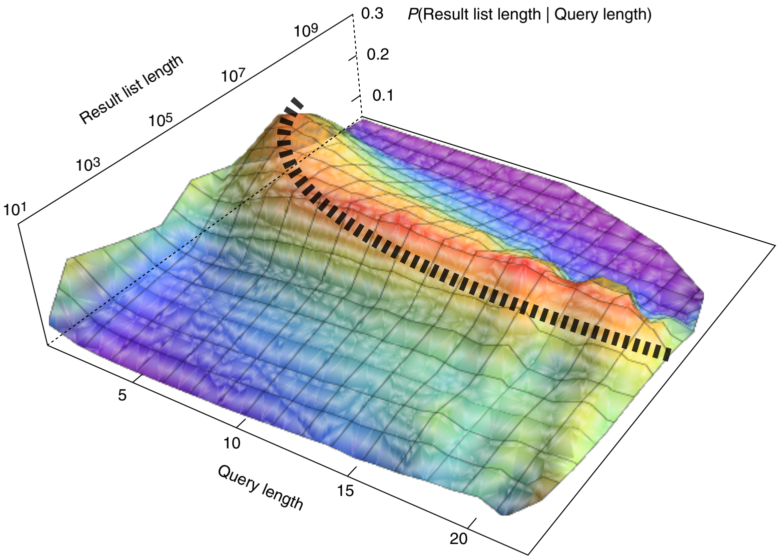
Probability for Retrieval Success

What about empirical evidence?

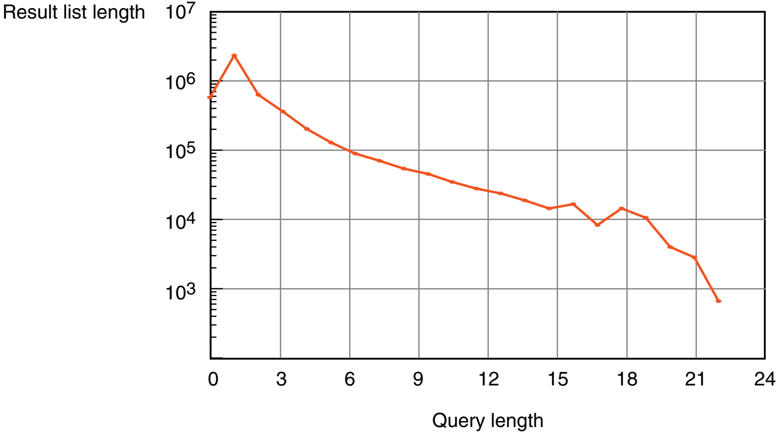
Experimental Setting: AOL log

- Cleaning (bots, URL queries, encoding problems)
- Query duplicates removed
- 4.4 million unique queries ($\leq$ 22 keywords)
- Submitted to Bing API
- Result list length estimates stored

AOL log result list length distribution in 3D



Median AOL log result list length in 2D

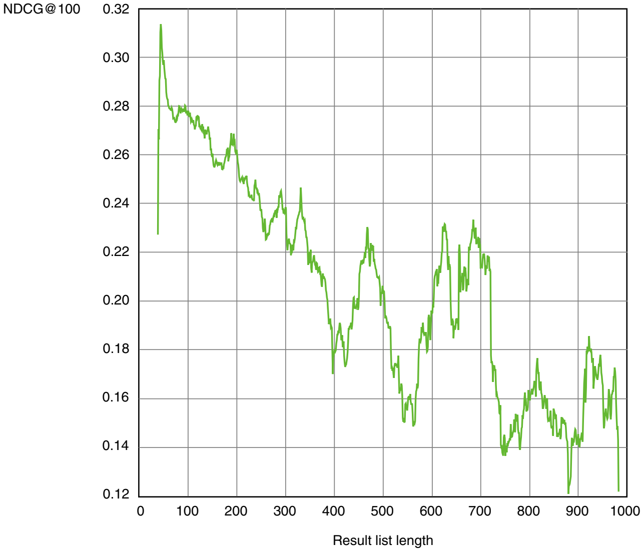


Experimental Setting: TREC Robust04

- 530 000 newswire documents
- BM25 indexed with Terrier
- Nounphrase extraction for TREC topics 301–450, 601–700
- Submitted all combinations to Terrier
- Result lists stored

- Assumed capacity $k = 100$

Avg. NDCG@100 per result list length (Robust04)



Almost the end: The take-away messages!

What we have done

Results

- When ranking works: fine!
- Else: User-over-Ranking
  - longer queries → fewer results
  - optimum retrieval performance → user capacity
- Empirical evidence

Future Work

- Apply hypothesis to query formulation

What we have (not) done

Results

- When ranking works: fine!
- Else: User-over-Ranking
  - longer queries → fewer results
  - optimum retrieval performance → user capacity
- Empirical evidence

Future Work

- Apply hypothesis to query formulation

What we have (not) done

Results

- When ranking works: fine!
- Else: User-over-Ranking
  - longer queries → fewer results
  - optimum retrieval performance → user capacity
- Empirical evidence

Future Work

- Apply hypothesis to query formulation

Thank you

☺