

Search Strategies for Keyword Queries

Matthias Hagen Benno Stein
Bauhaus-Universität Weimar
www.webis.de

Search Strategies for Keyword Queries

- ❑ Introduction
- ❑ Problem Statement
- ❑ Search Strategies
- ❑ Analysis and Results
- ❑ User over Ranking

Search Strategies for Keyword Queries

Introduction



"information retrieval"

Search

About **2,490,000** results (0.10 seconds)

[Advanced search](#)

Everything

Videos

Books

Discussions

Blogs

More

Any time

Past 2 days

All results

[Related searches](#)

[Wonder wheel](#)

[Timeline](#)

More search tools

[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](#)

Search Strategies for Keyword Queries

Introduction



"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 Problem, 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](#)

Search Strategies for Keyword Queries

Introduction



"information retrieval" "query formulation" "Web search"

Search

About **9,850** results (0.23 seconds)

[Advanced search](#)

Everything

More

All results

[Related searches](#)

[Wonder wheel](#)

[Page previews](#)

More search tools

[Scholarly articles for "information retrieval" "query formulation" "Web search"](#)



[Modern information retrieval](#) - Baeza-Yates - Cited by 7825

[Toward the semantic geospatial web](#) - Egenhofer - Cited by 251

[Information retrieval on the semantic web](#) - Shah - Cited by 142

[PDF] [QUERY FORMULATION IN WEB INFORMATION SEARCH](#)

File Format: PDF/Adobe Acrobat - [Quick View](#)

by A Aula - [Cited by 34](#) - [Related articles](#)

Query formulation is an essential part of successful **information retrieval**. The challenges in formulating effective queries ...

[citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.76... - Similar](#)

[PDF] [Download - User-Chosen Phrases in Interactive Query Formulation ...](#)

File Format: PDF/Adobe Acrobat - [Quick View](#)

by AF Smeaton - [Cited by 27](#) - [Related articles](#)

via a conventional **web search** engine. Recent work by Niwa et al. [13] has also presented an **Query Formulation** as an **Information Retrieval**. Problem. ...

[citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.52.9990&rep...](#)

[Show more results from citeseerx.ist.psu.edu](#)

Search Strategies for Keyword Queries

Introduction



"information retrieval" "search session"

Search

About 5,920 results (0.16 seconds)

[Advanced search](#)

Everything

More

All results

[Related searches](#)

[Wonder wheel](#)

[Page previews](#)

More search tools

[Scholarly articles for "information retrieval" "search session"](#)



[... -sensitive information retrieval using implicit feedback](#) - Shen - Cited by 175

[... online monitoring methods for information retrieval ...](#) - Borgman - Cited by 72

[Improving web search ranking by incorporating user ...](#) - Agichtein - Cited by 285

[An Overview of the Z39.50 Information Retrieval Standard - UDT ...](#)

by F Turner - Cited by 10 - [Related articles](#)

Z39.50 is an American national standard for **information retrieval**. ... to the searcher, keeping track of the results, terminating a **search session**, etc. ...

www.ifla.org/VI/5/op/udtop3/udtop3.htm - [Cached](#) - [Similar](#)

[Exploiting Session Context for Information Retrieval - A ...](#)

by G Pandey - 2008 - Cited by 1 - [Related articles](#)

of the current **search session**. In this work, we present a comparative tion for **information retrieval**. In: SIGIR 2001 (2001) ...

www.springerlink.com/index/200p0r260383u680.pdf

[\[PDF\] A Session-Based Search Engine](#)

File Format: PDF/Adobe Acrobat - [Quick View](#)

by S Sriram - Cited by 23 - [Related articles](#)

of clicked web pages) in the same **search session** and the session ... **information retrieval** toolkit. We design and implement a session-based search engine ...

citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.1101&rep...

Search Strategies for Keyword Queries

Introduction



"search session" "user support"

Search

About 344 results (0.11 seconds)

[Advanced search](#)

Everything

More

Show search tools

[\[PDF\] Search histories for user support in user interfaces](#)

File Format: PDF/Adobe Acrobat - [Quick View](#)

by A Komlodi - 2006 - [Cited by 23](#) - [Related articles](#)

users by visualizing **search session** histories. The system ... for **user support**.

Methodology. The project began with a field study of 16 attorneys and ...

citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.1649&rep...

[\[PDF\] The Re:Search Engine: Simultaneous Support for Finding and Re-Finding](#)

File Format: PDF/Adobe Acrobat - [Quick View](#)

by J Teevan - [Cited by 11](#) - [Related articles](#)

middle of a **search session**, it is likely that when a user is- histories for **user**

support in user interfaces. JASIST, 57(6): 803-807. ...

citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.8687...

[Search history support for finding and using information: User ...](#)

by A Komlodi - 2007 - [Cited by 10](#) - [Related articles](#)

After the search, they were interviewed about the **search session** and about their

..... Search history for **user support** in information-seeking interfaces. ...

linkinghub.elsevier.com/retrieve/pii/S0306457306000902

Search Strategies for Keyword Queries

Introduction



"search engine" "cost optimization"

Search

About 4,750 results (0.13 seconds)

[Advanced search](#)

Everything

More

All results

[Related searches](#)

[Wonder wheel](#)

[Timeline](#)

More search tools

[WikiAnswers - What is cost optimization](#)

Business Plans question: What is **cost optimization**? ... The cost of **search engine** optimization depends on what seo services you are looking at. ...

[wiki.answers.com/Q/What_is_cost_optimization](#) - [Cached](#) - [Similar](#)

[Search Engine Optimization | Links page](#)

Low **cost Optimization**. Top Google Rankings. All Major Search Engines. Proven Results. Top Ten Listings. Google Friendly Methods. **Search Engine** Marketing ...

[www.deeho.co.uk/links16.shtml](#) - [Cached](#)

[Search Engine Optimization Western Cape](#)

20 Aug 2010 ... **Search Engine** Optimization is actually much harder than it looks at the ... Construction Scheduling, **Cost Optimization** and Management ...

[www.docstoc.com/docs/.../Search-Engine-Optimization-Western-Cape](#)

[How Much Does It Cost?: Optimization of Costs in Sequence Analysis ...](#)

How Much Does It **Cost?: Optimization** of Costs in Sequence Analysis of Social Science Data. ... Pubget is a **search engine** that gets science PDFs fast. ...

[pubget.com/search?q=How+Much+Does+It...of...](#) - [Cached](#)

Search Strategies for Keyword Queries

Introduction

The complete session:

1. “information retrieval”
2. “information retrieval” “query formulation”
3. “information retrieval” “query formulation” “Web search”
4. “information retrieval” “search session”
5. “search session” “user support”
6. “search engine” “cost optimization”

Search Strategies for Keyword Queries

Introduction

The complete session:

1. "information retrieval"
2. "information retrieval" "query formulation"
3. "information retrieval" "query formulation" "Web search"
4. "information retrieval" "search session"
5. "search session" "user support"
6. "search engine" "cost optimization"

The $\bigcup_{i \in 1 \dots 6} \{\text{query}_i\}$ query:



"information retrieval" "query formulation" "Web search" "search ses: Search

[Advanced search](#)

Everything

More

All results

[Related searches](#)

[Wonder wheel](#)

[Page previews](#)

More search tools

Your search - "information retrieval" "query formulation" "Web search" "search session" "user support ... - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Search Strategies for Keyword Queries

Introduction

The complete session:

1. “information retrieval”
2. “information retrieval” “query formulation”
3. “information retrieval” “query formulation” “Web search”
4. “information retrieval” “search session”
5. “search session” “user support”
6. “search engine” “cost optimization”

The $\bigcup_{i \in 1 \dots 6} \{\text{query}_i\} \setminus \{\text{“user support”}\}$ query:



“information retrieval” “query formulation” “Web search” “search ses: Search

1 result (0.22 seconds)

[Advanced search](#)

Everything

More

All results

[Related searches](#)

[Wonder wheel](#)

[\[PDF\] Making the Most of a Web Search Session](#)

File Format: PDF/Adobe Acrobat - [View as HTML](#)

by B Stein

Keywords-**Web Search Session, Query Formulation, Query. Cost Optimization**

..... MedSearch: a specialized **search engine** for medical **information retrieval**.

...

www.uni-weimar.de/medien/webis/publications/.../papers/stein_2010n.pdf

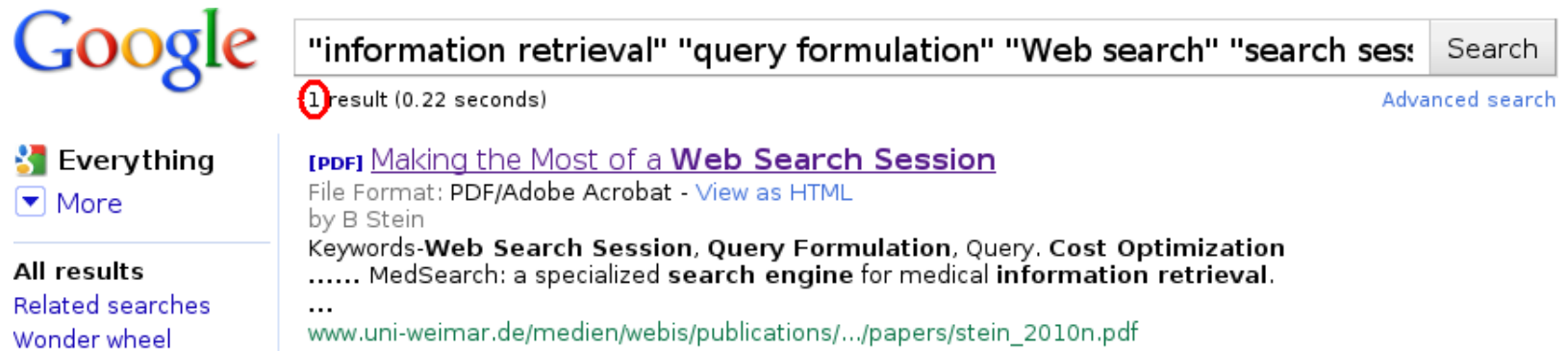
Search Strategies for Keyword Queries

Introduction

The complete session:

1. “information retrieval”
2. “information retrieval” “query formulation”
3. “information retrieval” “query formulation” “Web search”
4. “information retrieval” “search session”
5. “search session” “user support”
6. “search engine” “cost optimization”

The $\bigcup_{i \in 1 \dots 6} \{\text{query}_i\} \setminus \{\text{“user support”}\}$ query:



The screenshot shows a Google search interface. The search bar contains the query: "information retrieval" "query formulation" "Web search" "search ses". The search button is labeled "Search". Below the search bar, it indicates "1 result (0.22 seconds)" and a link to "Advanced search". The search results list a PDF document titled "[PDF] Making the Most of a Web Search Session" by B Stein. The keywords listed are "Web Search Session, Query Formulation, Query, Cost Optimization". The snippet of the document text reads: "..... MedSearch: a specialized search engine for medical information retrieval. ...". The URL of the document is "www.uni-weimar.de/medien/webis/publications/.../papers/stein_2010n.pdf". On the left side of the search results, there are links for "Everything", "More", "All results", "Related searches", and "Wonder wheel".

The maximum query.

Search Strategies for Keyword Queries

Problem Statement Maximum Query

Maximum query:

- An “as many keywords as possible” query
- Best single query to capture user’s articulated information need
- Ideally not too many results: user can check complete result list
 - “User over Ranking”
- Potential of improved user experience in search sessions

Search Strategies for Keyword Queries

Problem Statement Maximum Query

Maximum query:

- ❑ An “as many keywords as possible” query
- ❑ Best single query to capture user’s articulated information need
- ❑ Ideally not too many results: user can check complete result list
 - “User over Ranking”
- ❑ Potential of improved user experience in search sessions

Remarks:

- ❑ Maximum queries can be computed at server site or at client site.
- ❑ Current search engines do not suggest maximum queries.
- ❑ Analysis at client site ~ “Black-box index analysis”

Search Strategies for Keyword Queries

Problem Statement Maximum Query

Given:

1. A set W of keywords.
2. A query interface for a Web search engine S .
3. An upper bound k on the result list length.

Todo:

- Find a maximum subset $Q \subseteq W$ yielding at most k Web results.

Consider cost: **Minimize the number of submitted Web queries to find Q .**

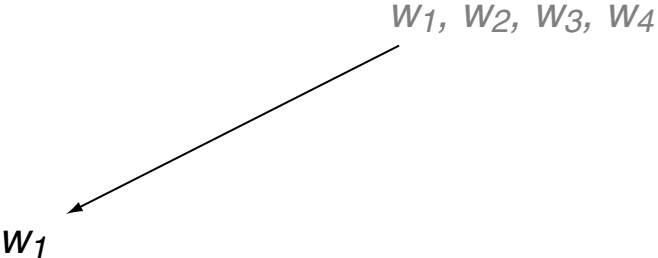
Search Strategies for Keyword Queries

Baseline Depth-First Search

W_1, W_2, W_3, W_4

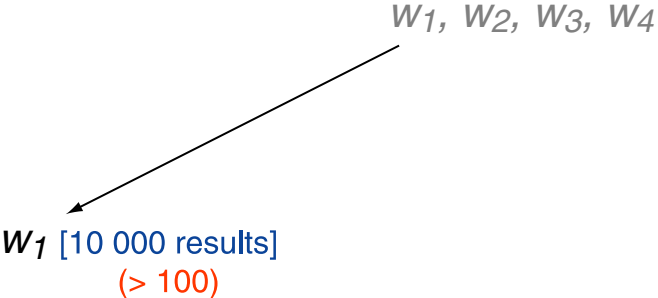
Search Strategies for Keyword Queries

Baseline Depth-First Search



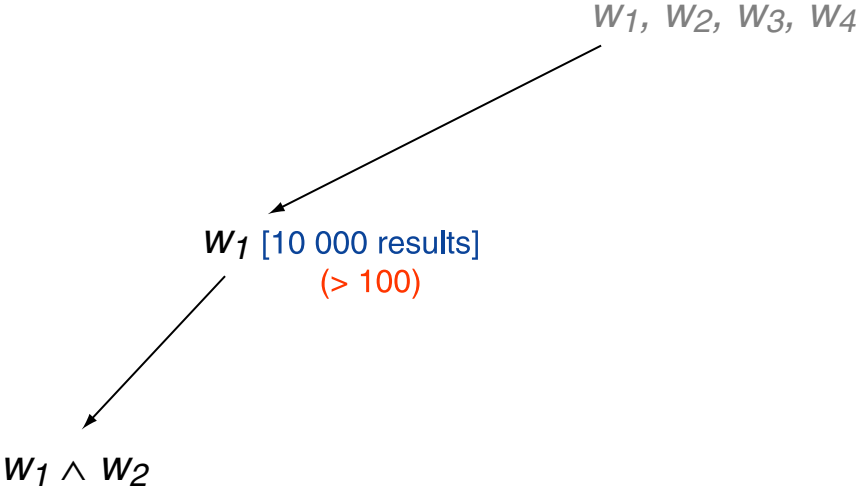
Search Strategies for Keyword Queries

Baseline Depth-First Search



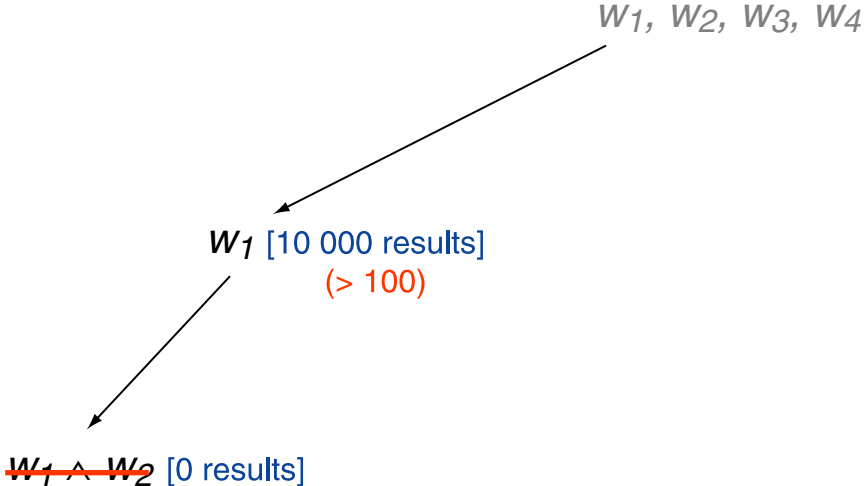
Search Strategies for Keyword Queries

Baseline Depth-First Search



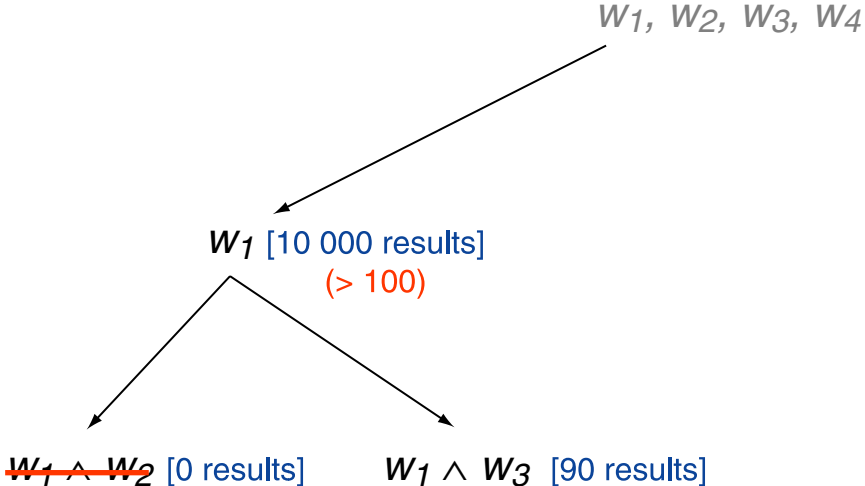
Search Strategies for Keyword Queries

Baseline Depth-First Search



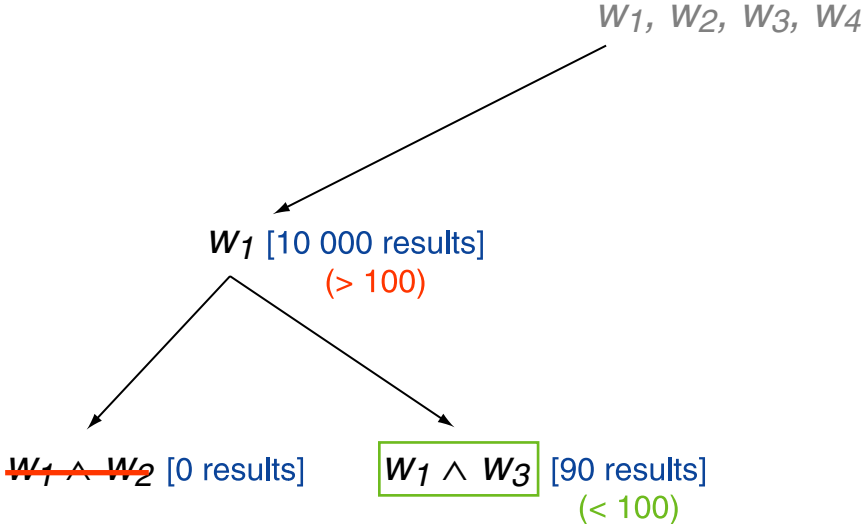
Search Strategies for Keyword Queries

Baseline Depth-First Search



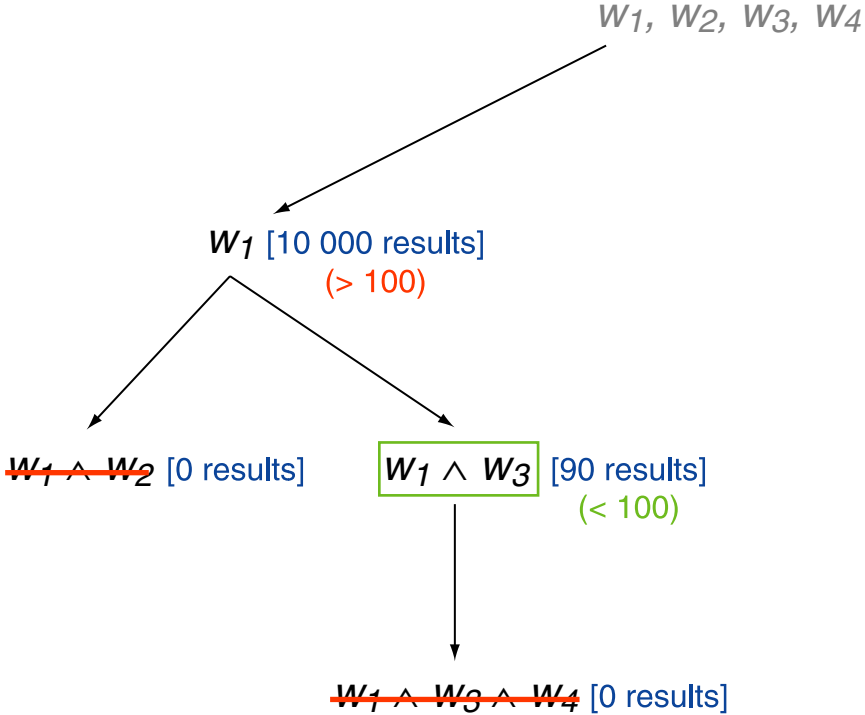
Search Strategies for Keyword Queries

Baseline Depth-First Search



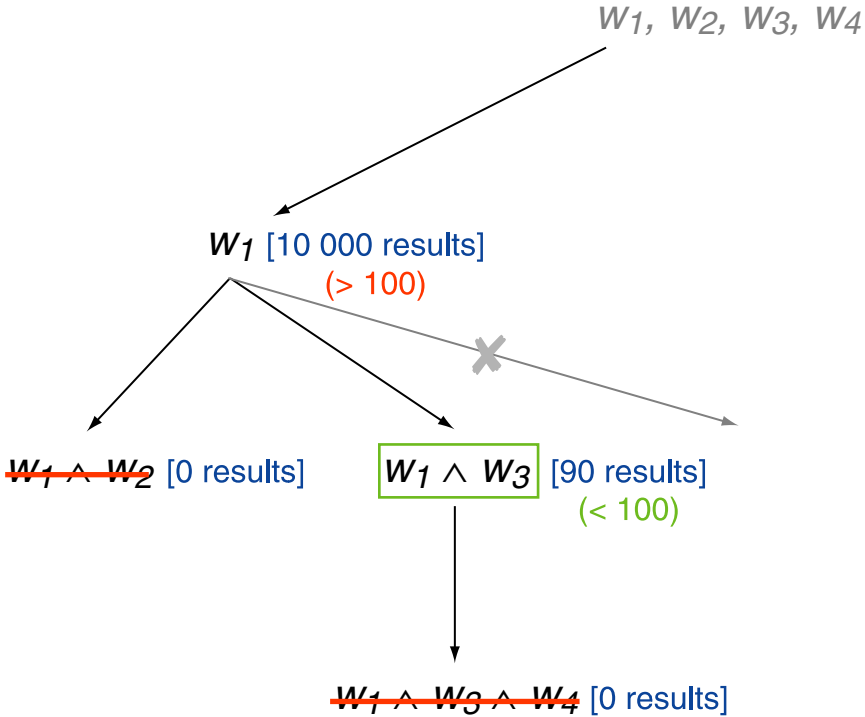
Search Strategies for Keyword Queries

Baseline Depth-First Search



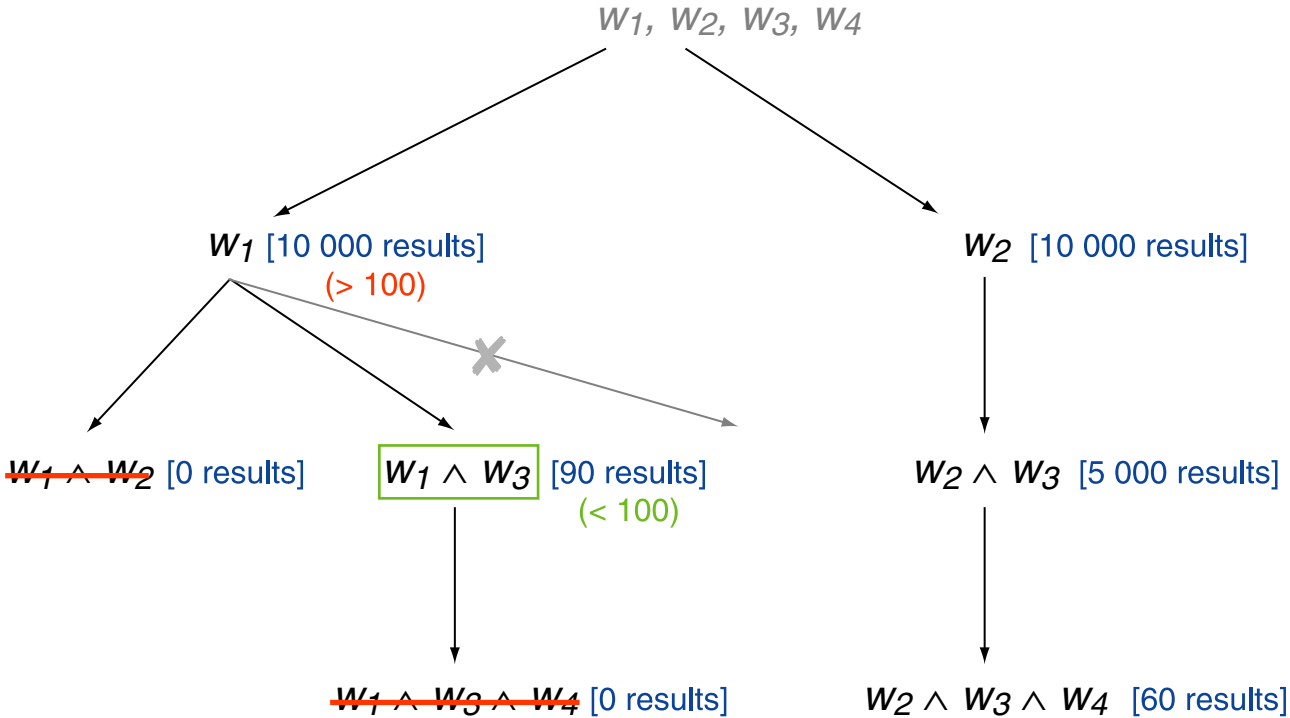
Search Strategies for Keyword Queries

Baseline Depth-First Search



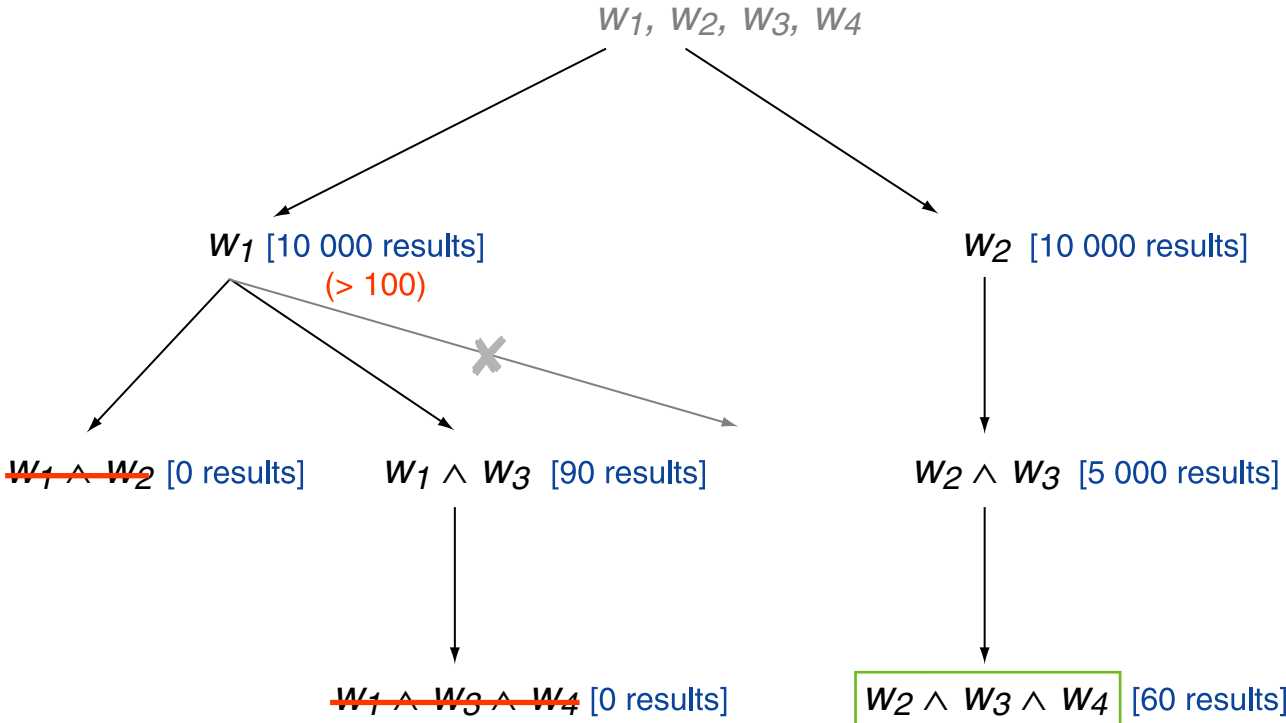
Search Strategies for Keyword Queries

Baseline Depth-First Search



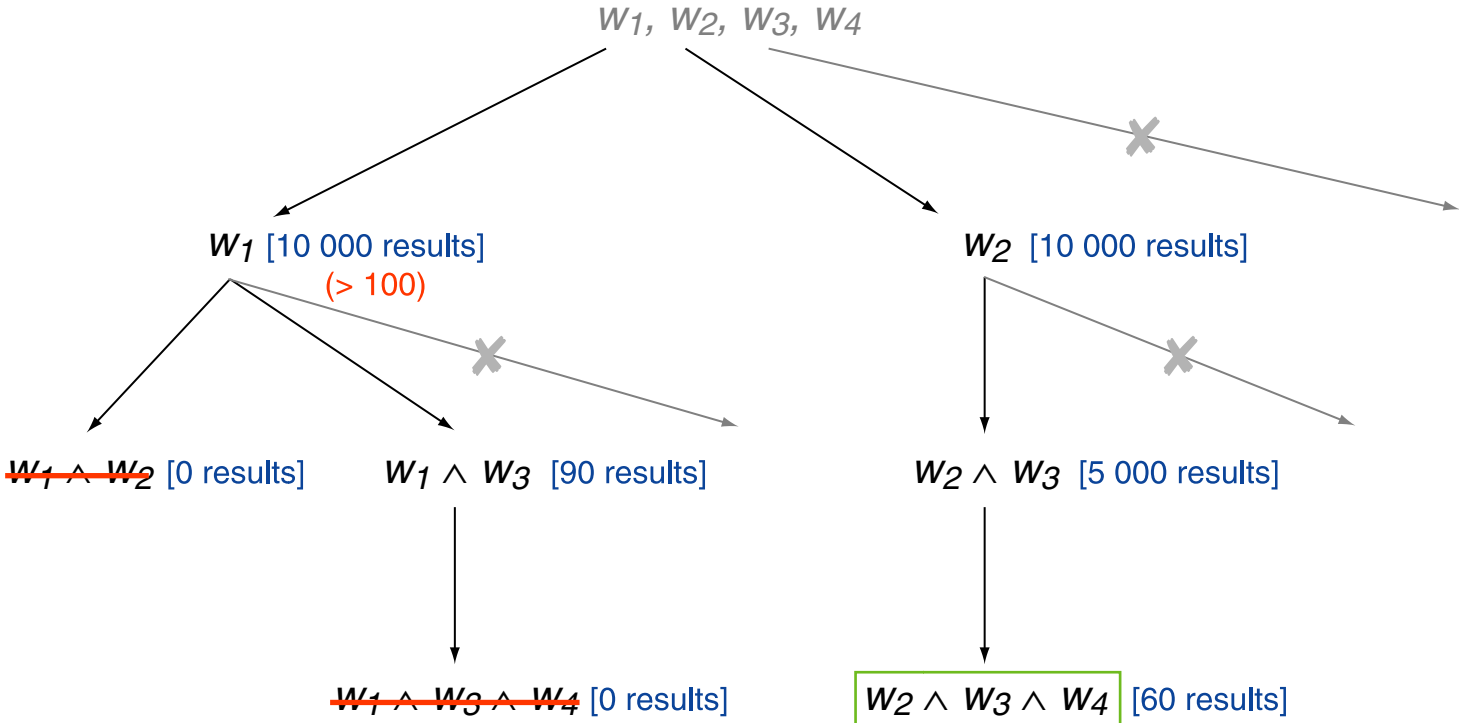
Search Strategies for Keyword Queries

Baseline Depth-First Search



Search Strategies for Keyword Queries

Baseline Depth-First Search



Search Strategies for Keyword Queries

Co-occurrence-based Web Count Estimation

Observation:

All intermediate queries are submitted to the search engine.

Concept of heuristic search:

Use under/over estimations for a query candidate's Web count before submission.

Search Strategies for Keyword Queries

Co-occurrence-based Web Count Estimation

Observation:

All intermediate queries are submitted to the search engine.

Concept of heuristic search:

Use under/over estimations for a query candidate's Web count before submission.

Example:

“information retrieval” 2,500,000 results (Google)

“information retrieval” “query formulation” 25,000 results (Google)

→ Yield factor $\gamma(\text{IR} + \text{QF}) = 0.01$

Search Strategies for Keyword Queries

Co-occurrence-based Web Count Estimation

Estimate:

“information retrieval” “query formulation” “Web search”

Numbers known so far:

“information retrieval” “query formulation” 25,000 results (Google)

$$\gamma(\text{IR} + \text{WS}) = 0.06$$

$$\gamma(\text{QF} + \text{WS}) = 0.16$$

Search Strategies for Keyword Queries

Co-occurrence-based Web Count Estimation

Estimate:

“information retrieval” “query formulation” “Web search”

Numbers known so far:

“information retrieval” “query formulation” 25,000 results (Google)

$$\gamma(\text{IR} + \text{WS}) = 0.06$$

$$\gamma(\text{QF} + \text{WS}) = 0.16$$

Our scheme: $25,000 \cdot \text{avg}(0.06, 0.16) = 2,750$ estimated

Search Strategies for Keyword Queries

Co-occurrence-based Web Count Estimation

Estimate:

“information retrieval” “query formulation” “Web search”

Numbers known so far:

“information retrieval” “query formulation” 25,000 results (Google)

$$\gamma(\text{IR} + \text{WS}) = 0.06$$

$$\gamma(\text{QF} + \text{WS}) = 0.16$$

Our scheme: $25,000 \cdot \text{avg}(0.06, 0.16) = 2,750$ estimated

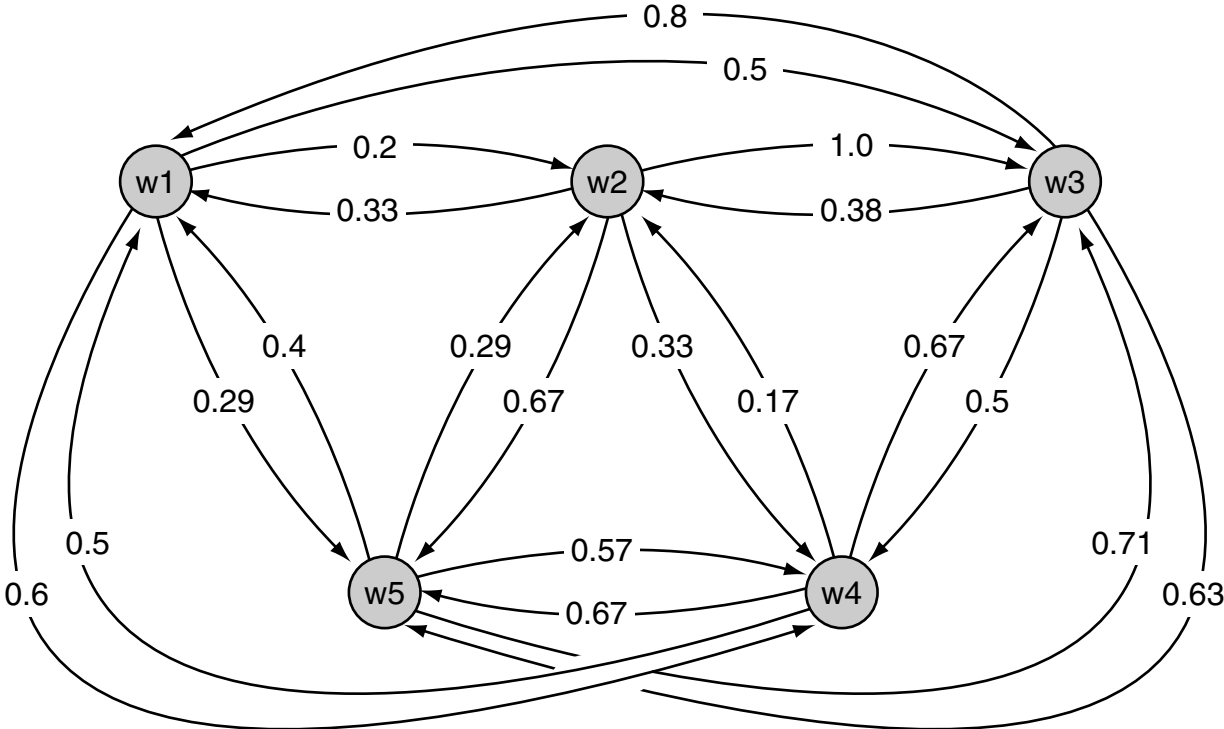
Comparison: 10,000 results (Google)

→ Co-occurrence-based estimations usually underestimate real Web count.

Search Strategies for Keyword Queries

Co-occurrence-based Web Count Estimation

During the search a co-occurrence graph is built up and maintained:



Remarks:

This graph can be built in a sandbox, e.g. based on Wikipedia.

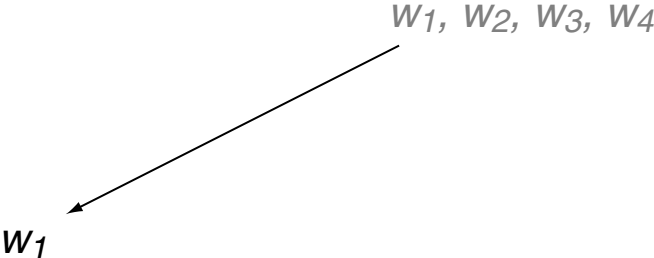
Search Strategies for Keyword Queries

Informed Search

W_1, W_2, W_3, W_4

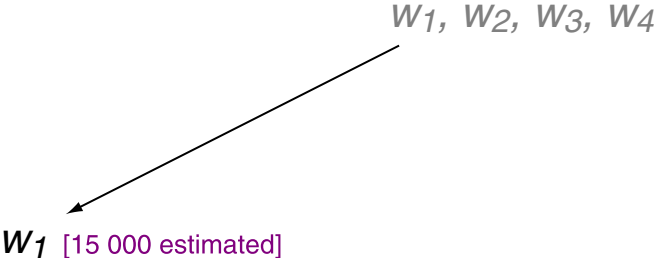
Search Strategies for Keyword Queries

Informed Search



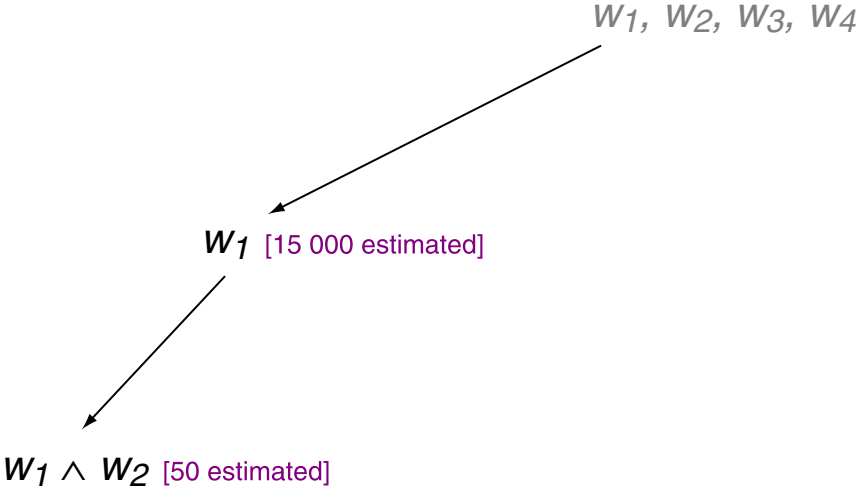
Search Strategies for Keyword Queries

Informed Search



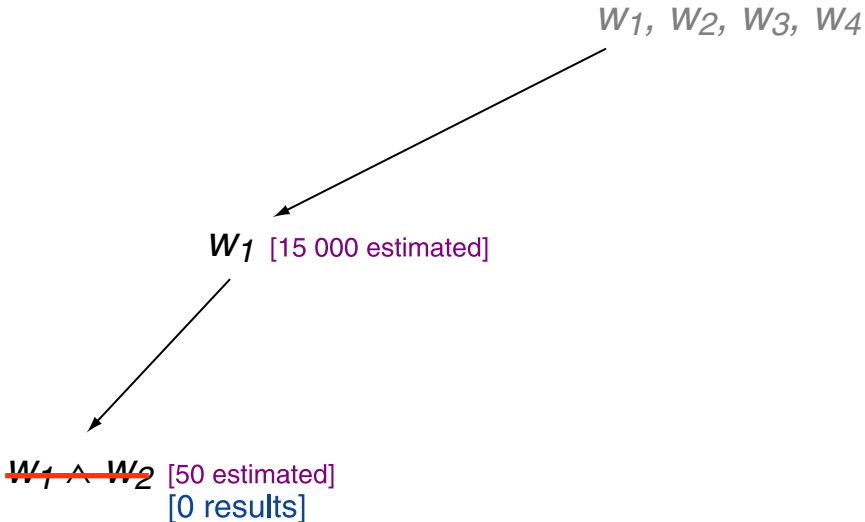
Search Strategies for Keyword Queries

Informed Search



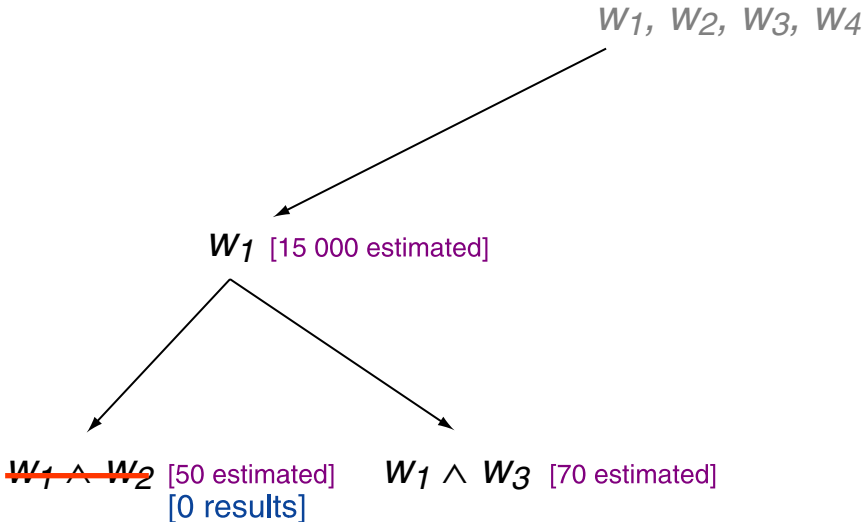
Search Strategies for Keyword Queries

Informed Search



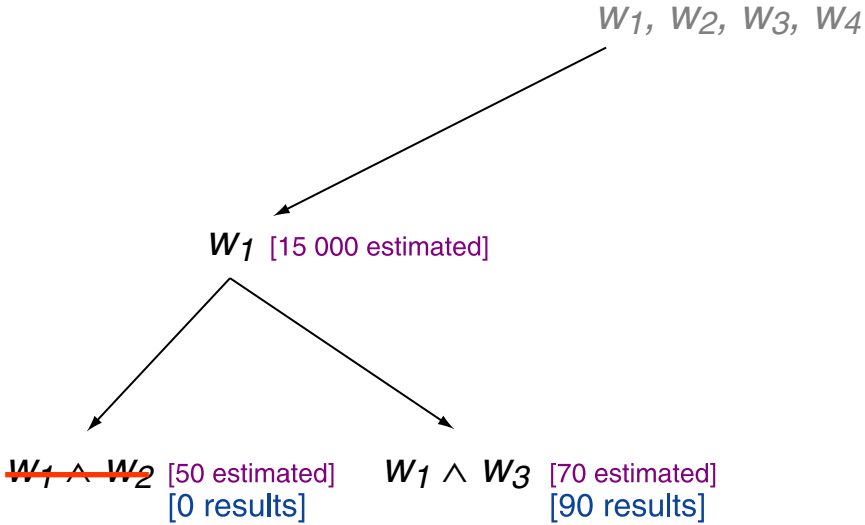
Search Strategies for Keyword Queries

Informed Search



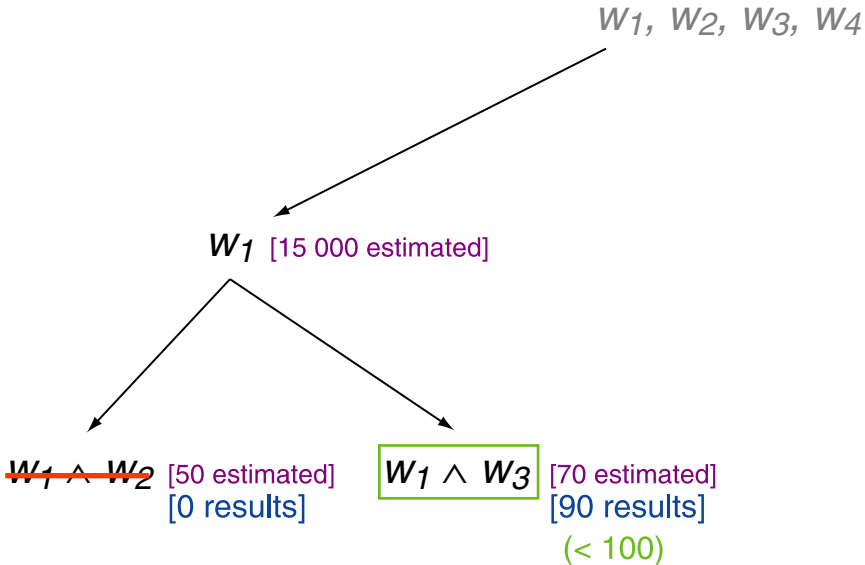
Search Strategies for Keyword Queries

Informed Search



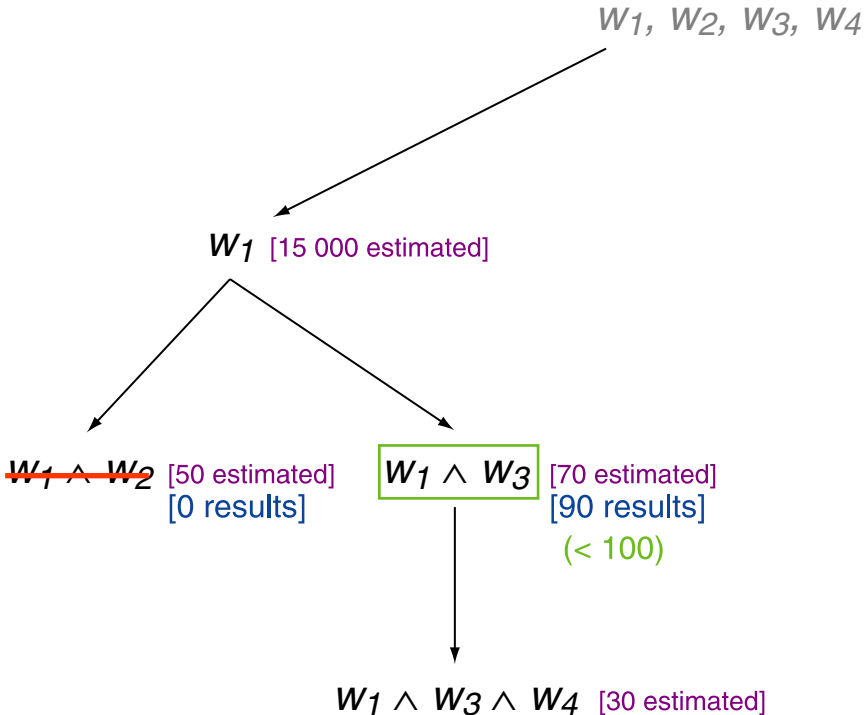
Search Strategies for Keyword Queries

Informed Search



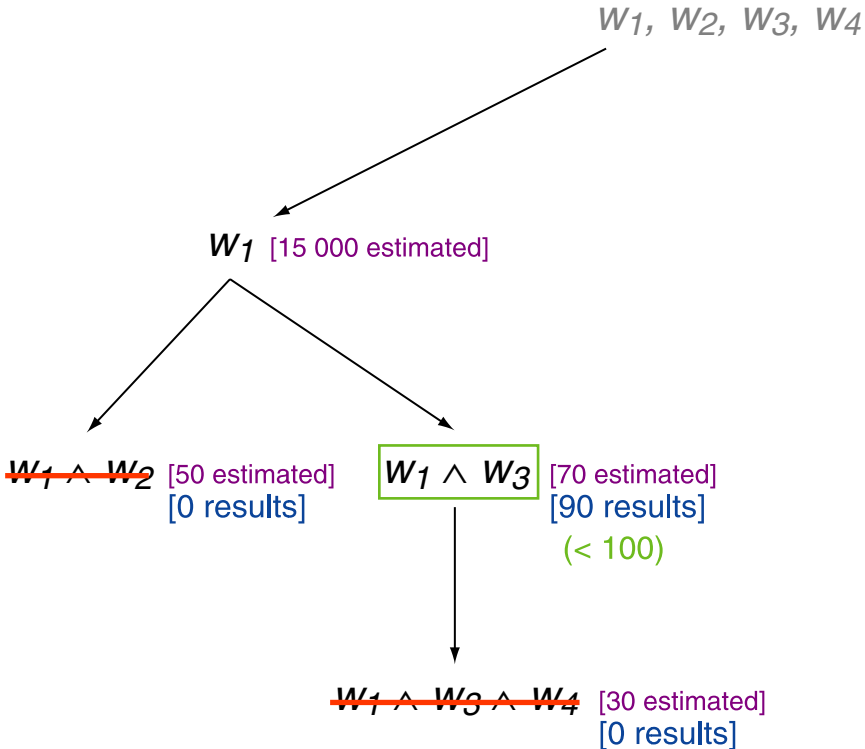
Search Strategies for Keyword Queries

Informed Search



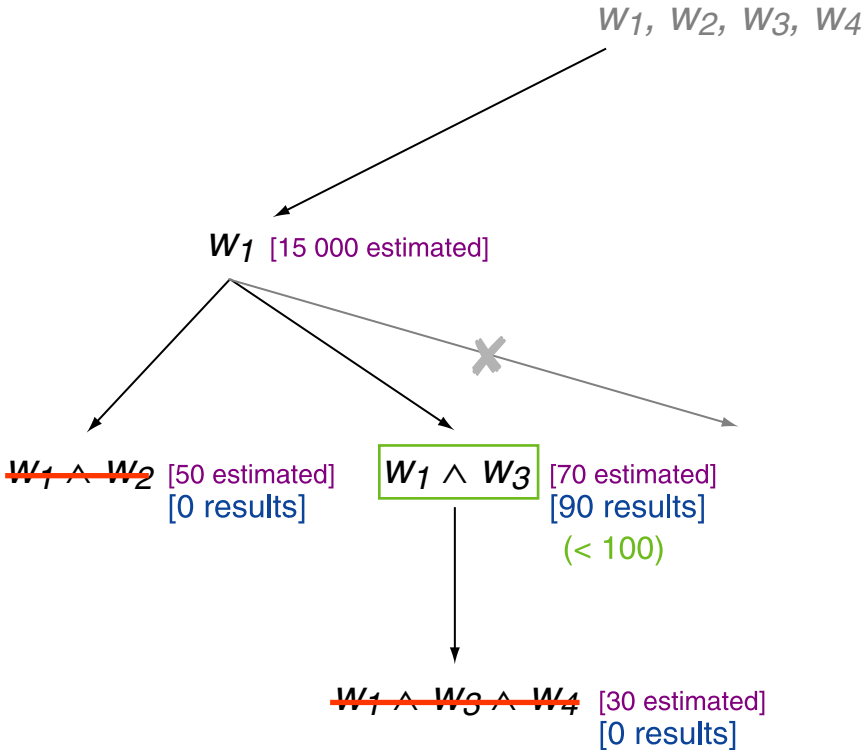
Search Strategies for Keyword Queries

Informed Search



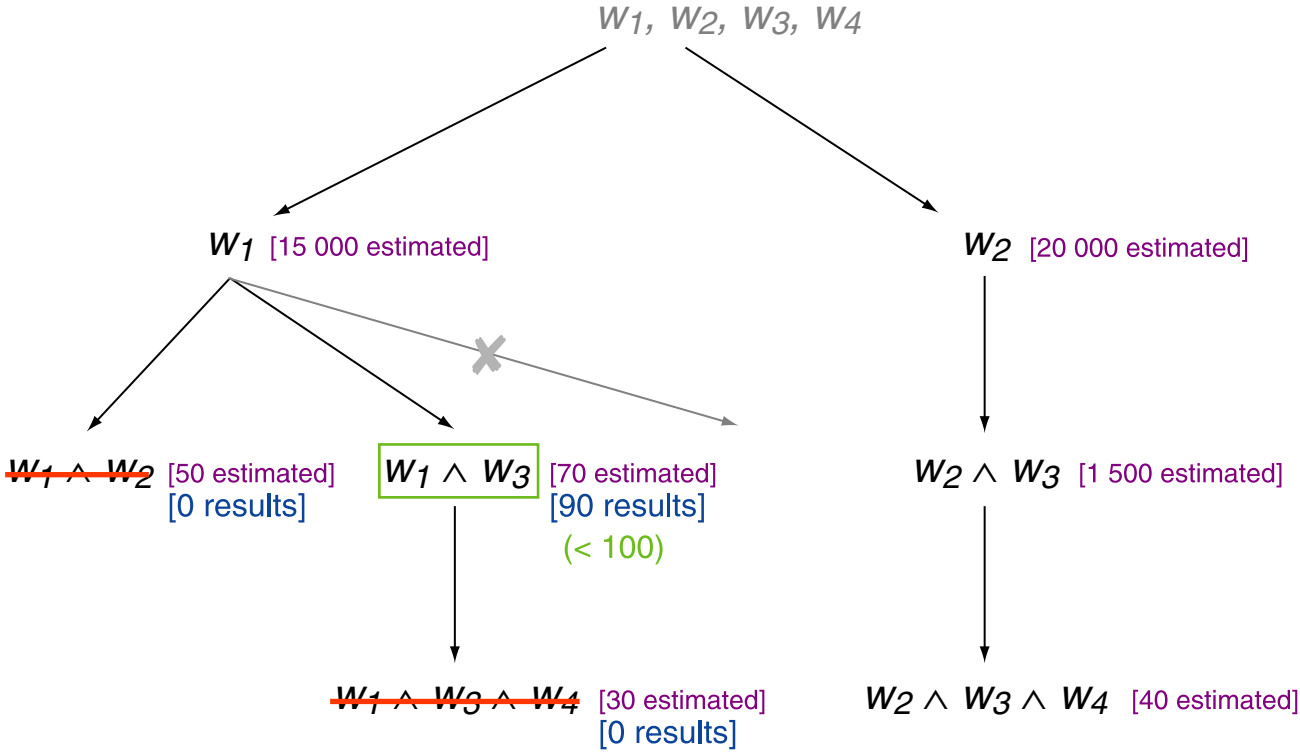
Search Strategies for Keyword Queries

Informed Search



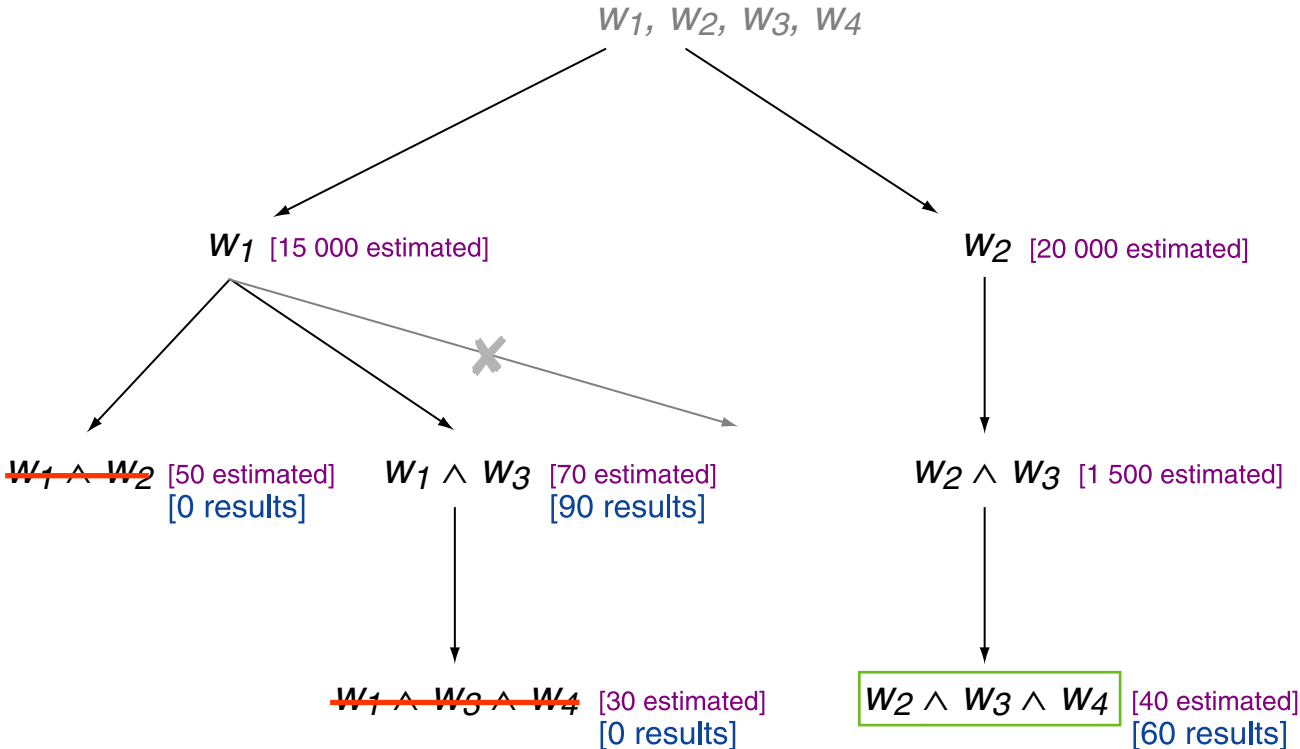
Search Strategies for Keyword Queries

Informed Search



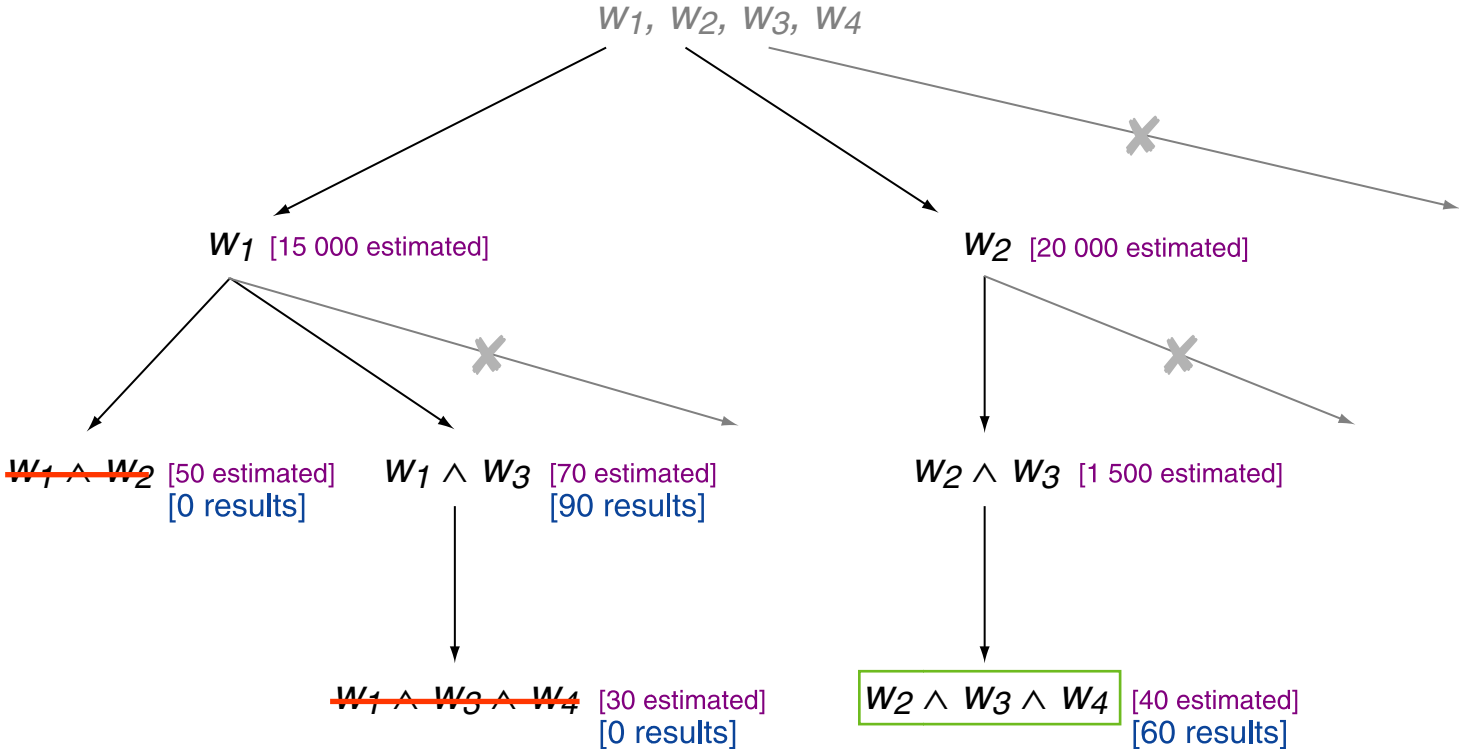
Search Strategies for Keyword Queries

Informed Search



Search Strategies for Keyword Queries

Informed Search



Search Strategies for Keyword Queries

Analysis and Results

Corpus and setup:

1. Collection with 775 CS papers from major conferences and journals.
2. 15 keywords are extracted per document, using extractor from [1].
3. Result set length $k = 100$ (\sim processing capacity).
4. Measure number of submitted Web queries (Bing API as search engine).

[1] Barker/Cornacchia. Using noun phrase heads to extract document keyphrases. Proc. AI 2000, pp. 40-52.

Search Strategies for Keyword Queries

Analysis and Results

Corpus and setup:

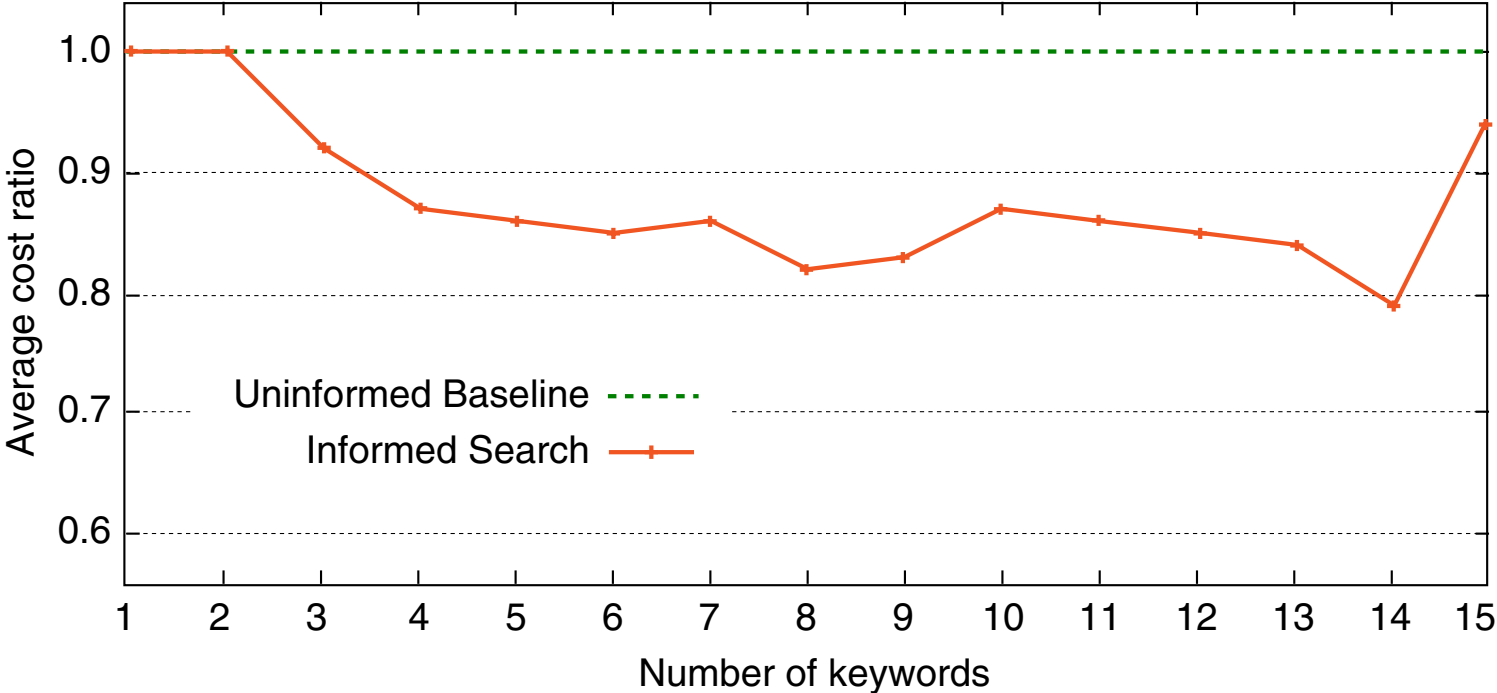
1. Collection with 775 CS papers from major conferences and journals.
2. 15 keywords are extracted per document, using extractor from [1].
3. Result set length $k = 100$ (\sim processing capacity).
4. Measure number of submitted Web queries (Bing API as search engine).

[1] Barker/Cornacchia. Using noun phrase heads to extract document keyphrases. Proc. AI 2000, pp. 40-52.

Number of keywords	5	10	15
No maximum query possible	595	328	86
Maximum query found	180	447	689
Avg. queries submitted informed	10.90	24.44	108.78
Avg. queries submitted baseline	12.67	29.40	116.22
Avg. Web query time (ms)	252.28	337.09	404.86
Avg. size maximum query informed	3.21	7.83	10.55
Avg. size maximum query baseline	3.21	7.90	10.57

Search Strategies for Keyword Queries

Analysis and Results



Search Strategies for Keyword Queries

Almost the End (The take-away messages ;-)

What we have done:

- Maximum Query problem statement
- External (client site) algorithms
- Co-occurrence based heuristics
- Heuristics outperform baselines
- Query Cover (in the WI paper)

Search Strategies for Keyword Queries

Almost the End (The take-away messages ;-)

What we have done:

- ❑ Maximum Query problem statement
- ❑ External (client site) algorithms
- ❑ Co-occurrence based heuristics
- ❑ Heuristics outperform baselines
- ❑ Query Cover (in the WI paper)

Open problems / work in progress:

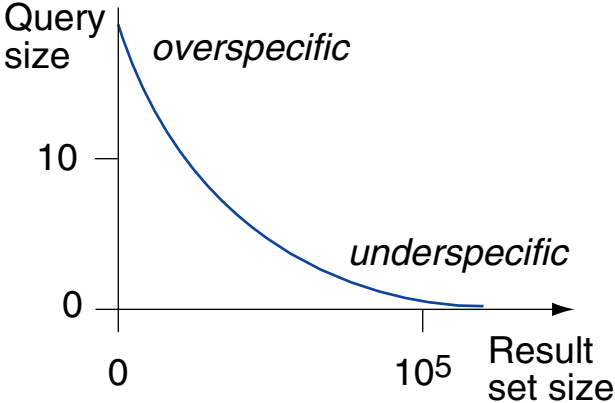
- ❑ Improved heuristics
- ❑ Co-occurrence source
- ❑ User study

Search Strategies for Keyword Queries

User over Ranking Hypothesis

Search Strategies for Keyword Queries

User over Ranking Hypothesis

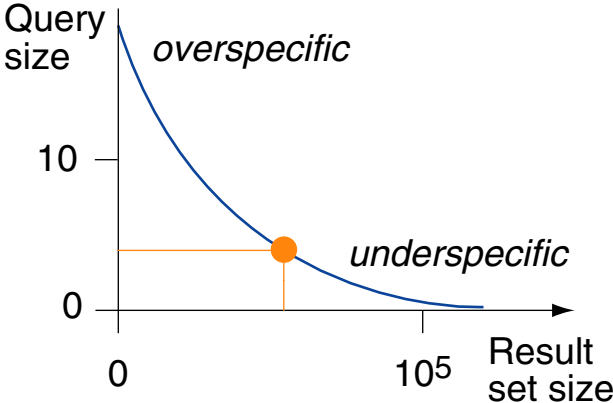


Query Specificity

Search Strategies for Keyword Queries

User over Ranking Hypothesis

- User can tell enough about her information need to overspecify a search.

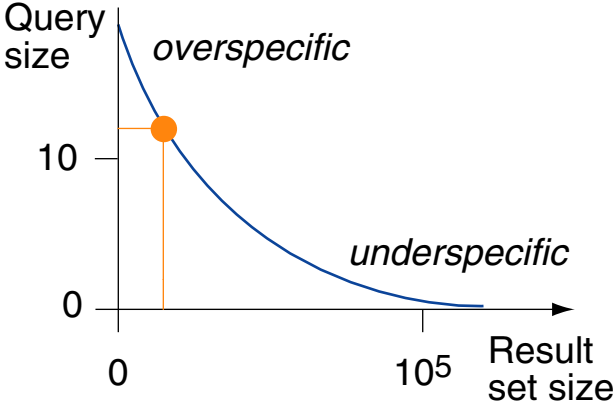


Query Specificity

Search Strategies for Keyword Queries

User over Ranking Hypothesis

- User can tell enough about her information need to overspecify a search.

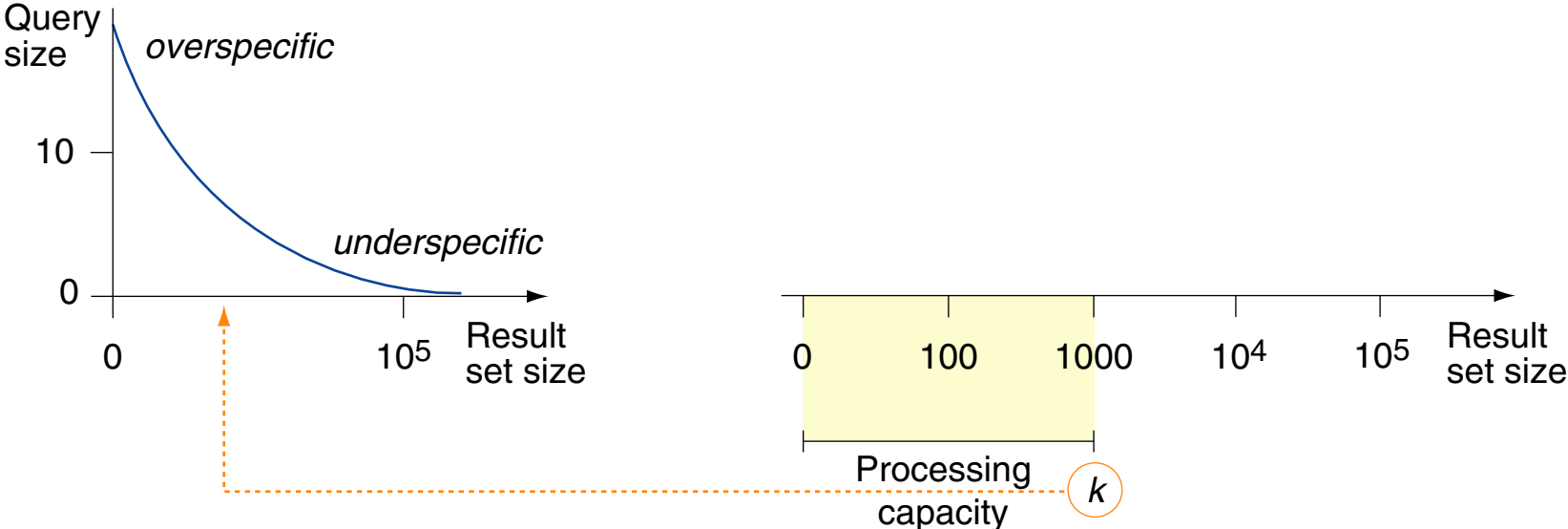


Query Specificity

Search Strategies for Keyword Queries

User over Ranking Hypothesis

- User can tell enough about her information need to overspecify a search.
- User can spent a certain amount of time to analyze results.

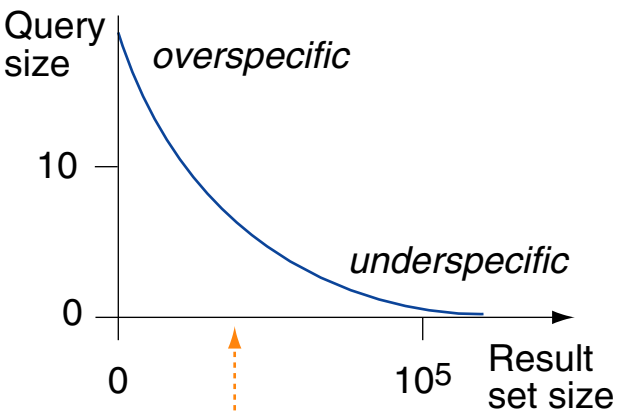


Query Specificity

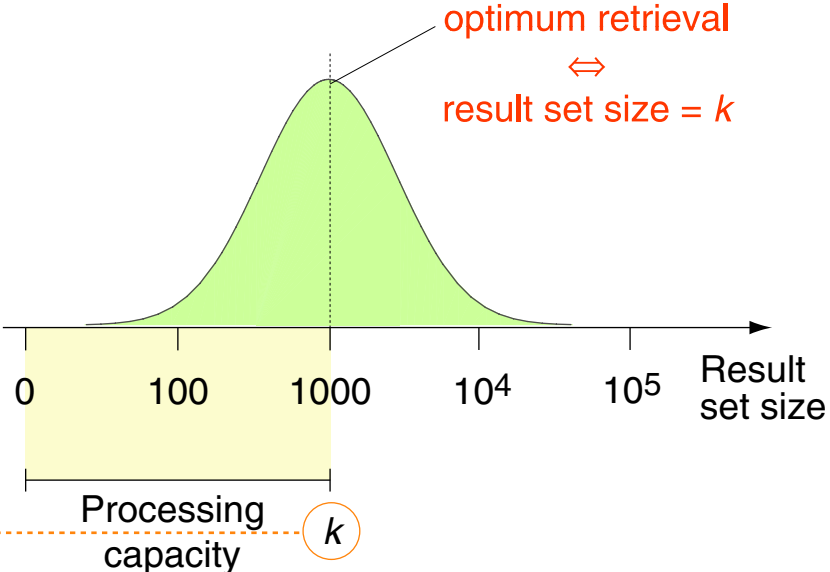
Search Strategies for Keyword Queries

User over Ranking Hypothesis

- User can tell enough about her information need to overspecify a search.
- User can spent a certain amount of time to analyze results.
- Rely on user rather than on ranking algorithms:
exploit processing capacity, considering “as many keywords as possible”.



Query Specificity



Probability for Retrieval Success

Thank you!