

# Making the most of a Web search session

Benno Stein    Matthias Hagen

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

Web Information and Quality Evaluation  
Valencia  
September 13, 2010

# Observing a sample user





 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](http://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](http://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](http://www.csli.stanford.edu/~hinrich/information-retrieval-book.html) - [Cached](#)
[Journal of Information Retrieval - SpringerLink Journal](#)
[www.springerlink.com/link.asp?id=103814](http://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](http://www.springer.com/computer/database+management.../10791) - [Cached](#)

# Observing a sample user – query 2



"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](#)

# Observing a sample user – query 3



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

Search

About 8,850 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](#)

# Observing a sample user – query 4





About 5,920 results (0.16 seconds)

[Advanced search](#)

Everything

**All results**
[Related searches](#)
[Wonder wheel](#)
[Page previews](#)


## 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.ifa.org/NI/5/op/udtop3/udtop3.htm](http://www.ifa.org/NI/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](http://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...](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.1101&rep...)

# Observing a sample user – query 5



"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...](http://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...](http://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](http://linkinghub.elsevier.com/retrieve/pii/S0306457306000902)

# Observing a sample user – query 6



"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](#)

# The complete search 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”



# How to get better results from the session?

All keywords as one query?

“information retrieval”      “query formulation”  
“Web search”      “search session”      “user support”  
“search engine”      “cost optimization”



"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.**

# How to get better results from the session?

Use **as many keywords as possible!**

“information retrieval”      “query formulation”  
 “Web search”      “search session”      ~~“user support”~~  
 “search engine”      “cost optimization”



"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](http://www.uni-weimar.de/medien/webis/publications/.../papers/stein_2010n.pdf)

# “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 list
- Potential of improved user experience in search sessions

# “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 list
- Potential of improved user experience in search sessions

## “Problem”

Current engines do not suggest “as many keywords as possible”-queries.  
So: How to find them?

# “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 list
- Potential of improved user experience in search sessions

## “Problem”

Current engines do not suggest “as many keywords as possible”-queries.  
So: How to find them?

## Idea:

Externally compute them at client site!

# As a Formal Problem Statement

## MAXIMUM QUERY

- Given:
  - ① A set  $W$  of keywords.
  - ② A query interface for a Web search engine  $\mathcal{S}$ .
  - ③ An upper bound  $k$  on the result list length.
- Find a maximum subset  $Q \subseteq W$  yielding at most  $k$  Web results.

## Optimization Problem!

Minimize the number of submitted Web queries to find  $Q$ .

# As a Formal Problem Statement

## MAXIMUM QUERY

- Given:
  - ① A set  $W$  of keywords.
  - ② A query interface for a Web search engine  $\mathcal{S}$ .
  - ③ An upper bound  $k$  on the result list length.
- Find a maximum subset  $Q \subseteq W$  yielding at most  $k$  Web results.

## Optimization Problem!

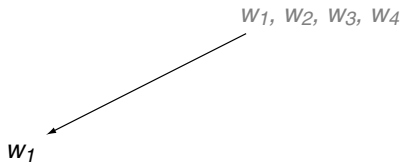
Minimize the number of submitted Web queries to find  $Q$ .

# Simple Attack: Depth-First Search

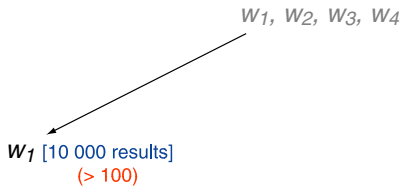
$W_1, W_2, W_3, W_4$



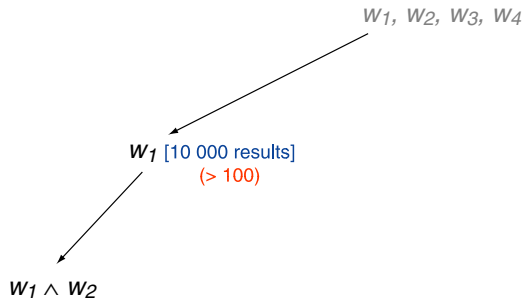
# Simple Attack: Depth-First Search



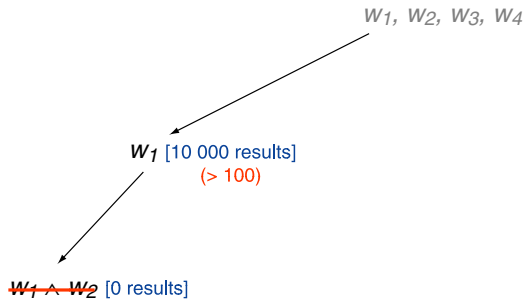
# Simple Attack: Depth-First Search



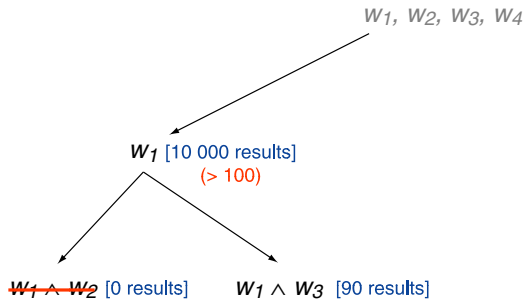
# Simple Attack: Depth-First Search



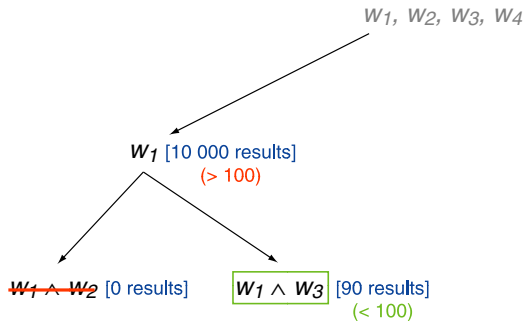
# Simple Attack: Depth-First Search



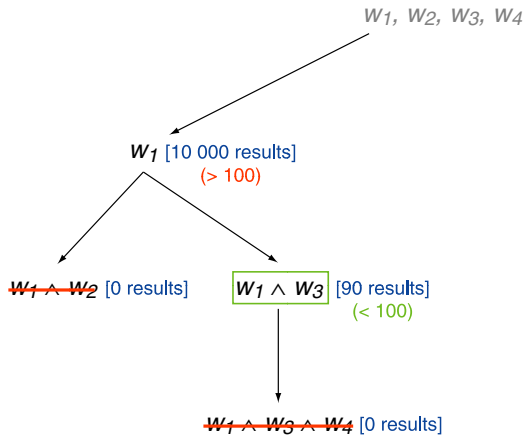
# Simple Attack: Depth-First Search



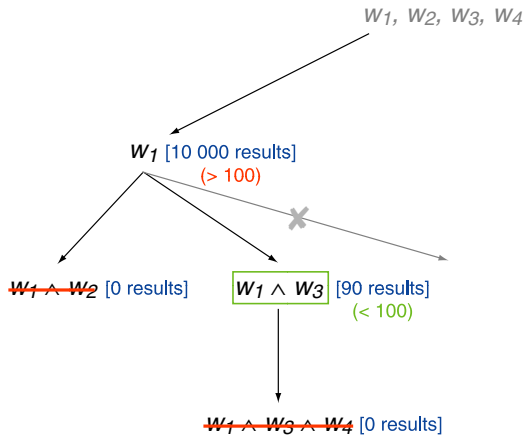
# Simple Attack: Depth-First Search



# Simple Attack: Depth-First Search

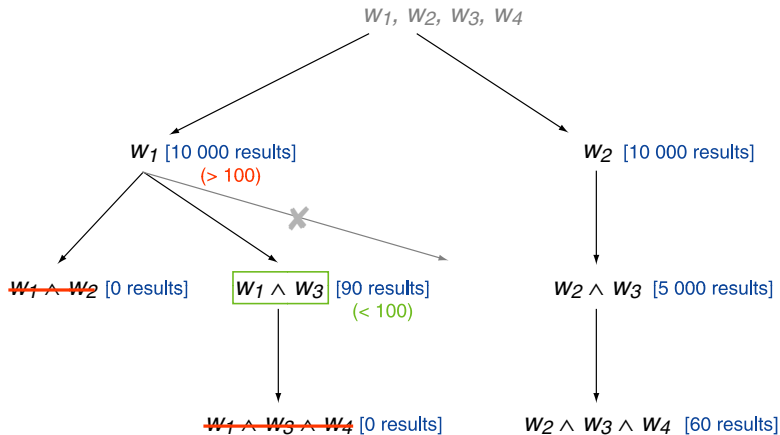


# Simple Attack: Depth-First Search

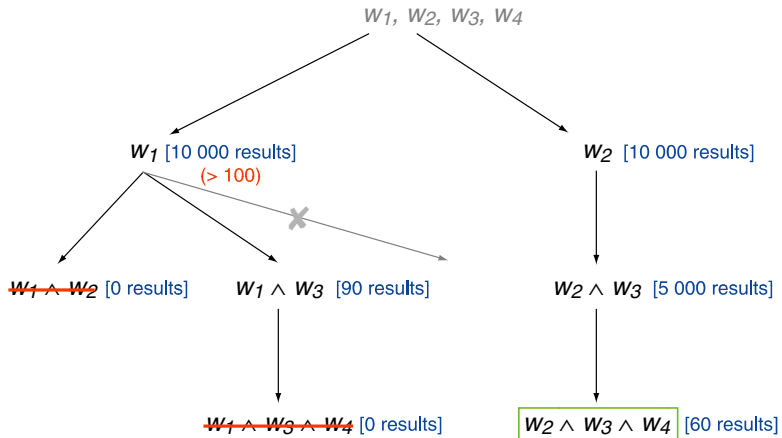




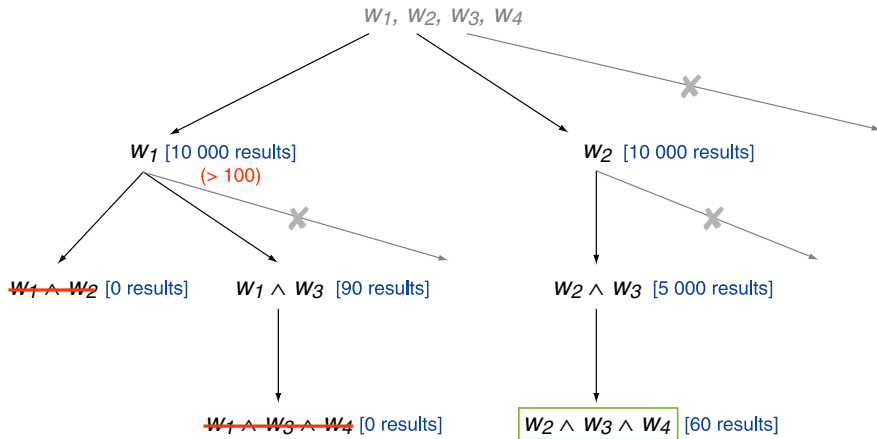
# Simple Attack: Depth-First Search



# Simple Attack: Depth-First Search



# Simple Attack: Depth-First Search



# Baseline's Analysis

## Major Drawback

All intermediate queries submitted to the search engine.  $\Rightarrow$  Bad run time!

# Baseline's Analysis

## Major Drawback

All intermediate queries submitted to the search engine.  $\Rightarrow$  Bad run time!

## Solution Idea:

Estimate a query candidate's Web count before submission.

# Co-occurrences

Google	“information retrieval”	2,500,000 results
Google	“information retrieval” “query formulation”	25,000 results

gives yield-factor:  $\gamma(\text{IR} + \text{QF}) = 0.01$

# Co-occurrence based Web Count Estimation

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

# Co-occurrence based Web Count Estimation

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

Known numbers:

25,000 results for “information retrieval” “query formulation”

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

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



# Co-occurrence based Web Count Estimation

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

Known numbers:

25,000 results for “information retrieval” “query formulation”

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

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

Our scheme:

$25,000 \cdot \text{avg}(0.06, 0.16) = 2,750$  results

# Co-occurrence based Web Count Estimation

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

Known numbers:

25,000 results for “information retrieval” “query formulation”

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

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

Our scheme:

$25,000 \cdot \text{avg}(0.06, 0.16) = 2,750$  results

Control:  10,000 results

# Co-occurrence based Web Count Estimation

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

Known numbers:

25,000 results for “information retrieval” “query formulation”

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

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

Our scheme:

$25,000 \cdot \text{avg}(0.06, 0.16) = 2,750$  results

Control:  10,000 results

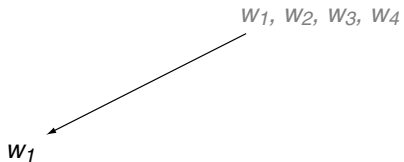
Observation:

Our co-occurrence based estimations usually underestimate real Web count.

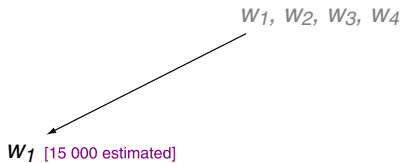
# “Informed” Depth-First Search

$W_1, W_2, W_3, W_4$

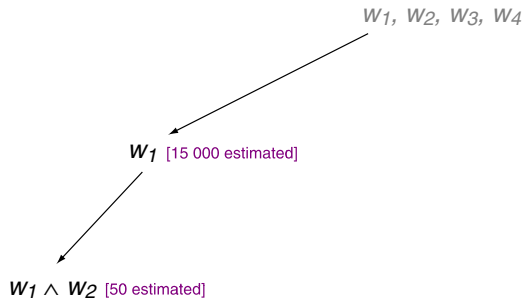
# “Informed” Depth-First Search



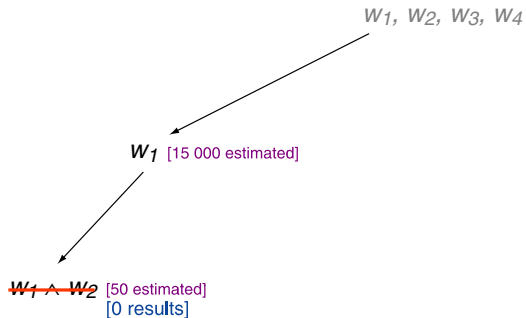
# “Informed” Depth-First Search



# “Informed” Depth-First Search

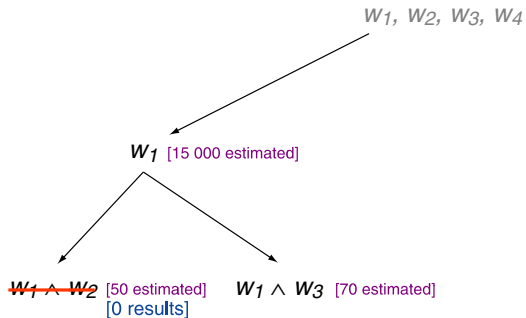


# “Informed” Depth-First Search

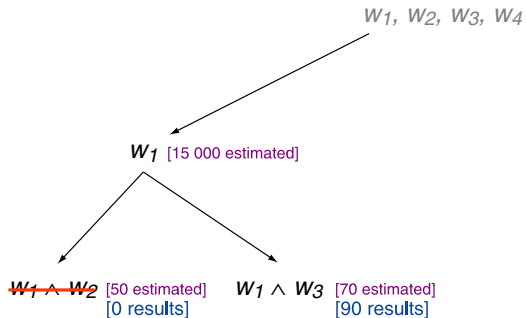




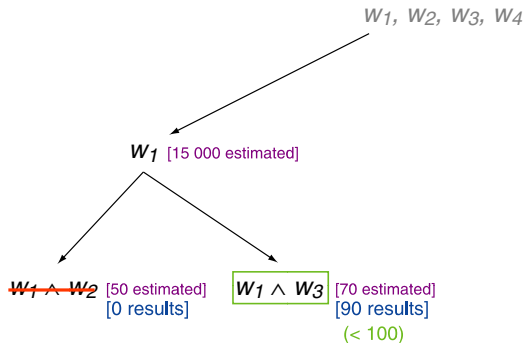
# “Informed” Depth-First Search



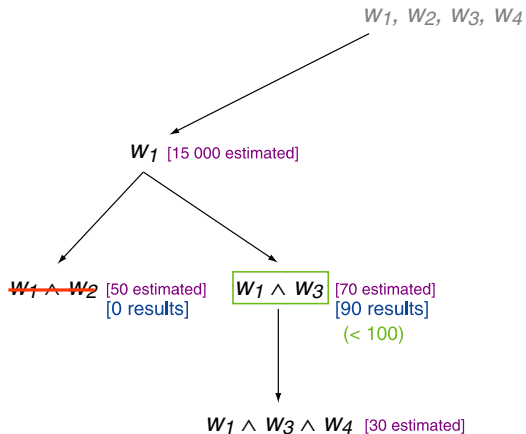
# “Informed” Depth-First Search



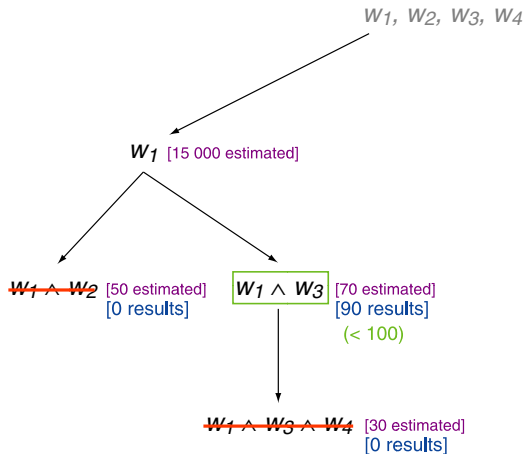
# “Informed” Depth-First Search



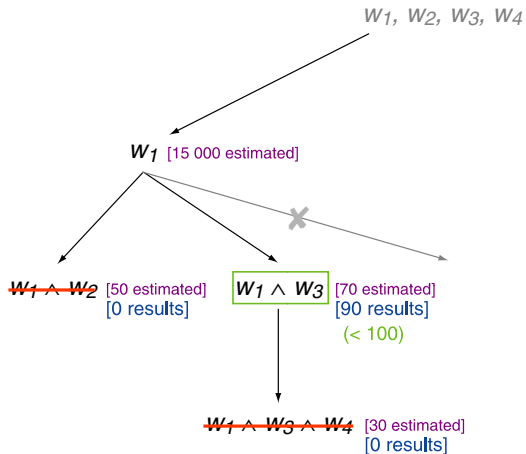
# “Informed” Depth-First Search



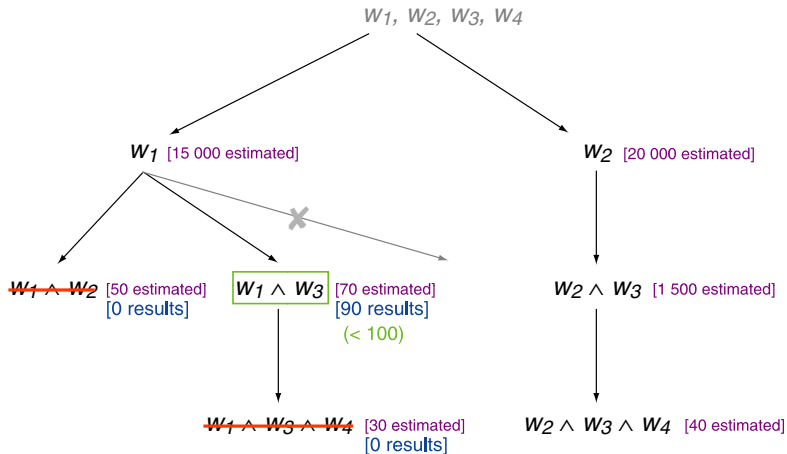
# “Informed” Depth-First Search



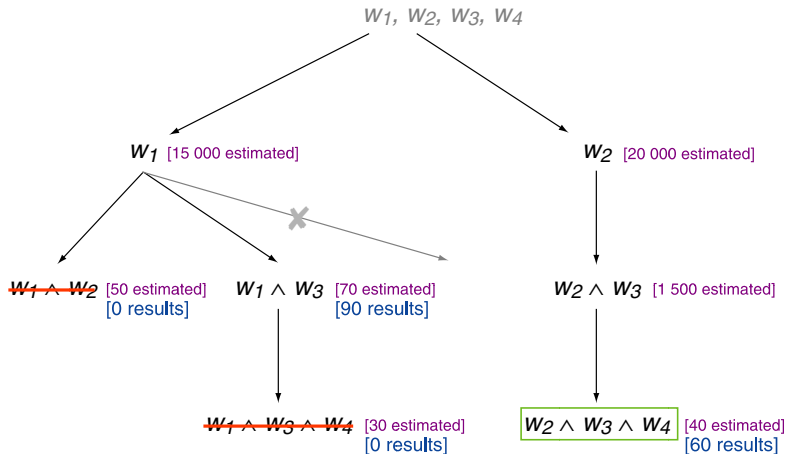
# “Informed” Depth-First Search



# “Informed” Depth-First Search

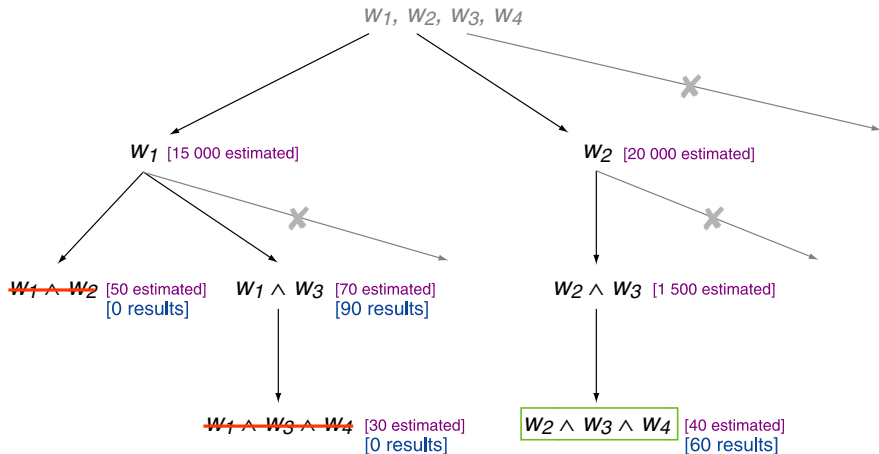


# “Informed” Depth-First Search





# "Informed" Depth-First Search



# Experimental Setup

## Corpus:

Extracted sessions ( $\geq 2$  queries) from AOL log with two methods:

- 10 minute time cut-off,
- Geometric method [Gayo-Avello 2009].

Removed stopwords.

For  $i \in \{3, \dots, 15\}$  sampled 1000  $i$ -keyword-sessions from each method.

## System:

Bing API as search engine.

Set  $k = 100$ .

Measured number of submitted Web queries.

# Experimental Setup

## Corpus:

Extracted sessions ( $\geq 2$  queries) from AOL log with two methods:

- 10 minute time cut-off,
- Geometric method [Gayo-Avello 2009].

Removed stopwords.

For  $i \in \{3, \dots, 15\}$  sampled 1000  $i$ -keyword-sessions from each method.

## System:

Bing API as search engine.

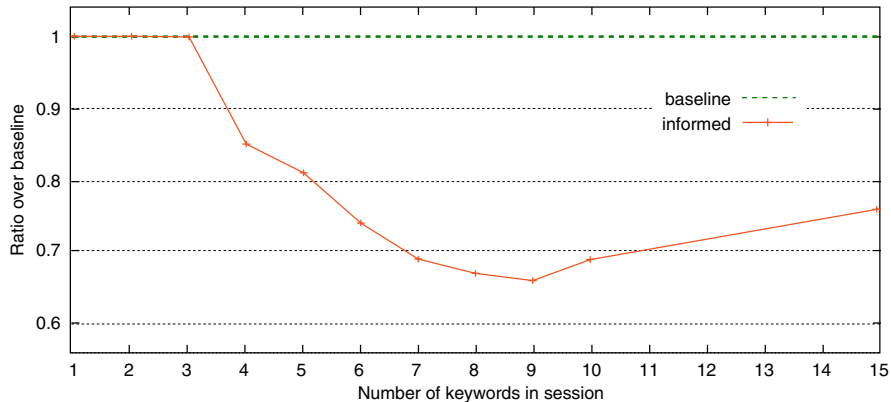
Set  $k = 100$ .

Measured number of submitted Web queries.

# Experimental Results

Number of keywords	5	10	15
No maximum query possible	1 913	1 599	953
Maximum query found	87	401	1 047
Avg. queries submitted informed	10.90	48.15	394.41
Avg. queries submitted baseline	13.38	70.29	516.46
Avg. Web query time (ms)	359.05	367.94	336.45
Avg. size maximum query informed	3.09	7.47	11.93
Avg. size maximum query baseline	3.19	7.71	12.34

# Experimental Results



Almost the end: The take-away messages!

# What we have done

## Results

- MAXIMUM QUERY
- QUERY COVER (in the paper)
- External (client site) algorithms
- Co-occurrence based heuristics
- Heuristics outperform baselines

## Open Problems

- Improved heuristics
- Co-occurrence source
- User study

# What we have (not) done

## Results

- MAXIMUM QUERY
- QUERY COVER (in the paper)
- External (client site) algorithms
- Co-occurrence based heuristics
- Heuristics outperform baselines

## Open Problems

- Improved heuristics
- Co-occurrence source
- User study



# What we have (not) done

## Results

- MAXIMUM QUERY
- QUERY COVER (in the paper)
- External (client site) algorithms
- Co-occurrence based heuristics
- Heuristics outperform baselines

## Open Problems

- Improved heuristics
- Co-occurrence source
- User study

**Thank you**  
