

Joint RuSSIR/EDBT Summer School 2011

WEB OF DATA

August 15-19 | 2011 | Saint Petersburg

EDBT RuSSIR



Caching query-biased snippets for efficient retrieval

Salvatore Orlando⁺, Raffaele Perego^{*}, Fabrizio Silvestri^{*}

^{*}ISTI - CNR, Pisa, Italy

⁺Università Ca' Foscari Venezia, Italy

Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri: **Caching query-biased snippets for efficient retrieval. EDBT 2011: 93-104**

Search Engine Results Page

Web Images Video Local Shopping News More ▾

edbt2011

▼

Search: ☒ the Web ☐ only in UK ☐ only in Ireland

Top-k results relevant for the query

TITLE

SNIPPET

URL

[EDBT/ICDT 2011](#)

March 21-25, 2011 --- Uppsala, Sweden. EDBT: 14th International Conference on Extending Database Technology, March 22-24 The EDBT series of conferences is an established ...

[edbticdt2011.it.uu.se](#) - [Cached](#)

[Dingming Wu](#)

I was visiting The Hong Kong Polytechnic University collaborating with Man Lung Yiu from March 1st to August 31st in 2010. I was an external reviewer of **EDBT2011**, VLDB J ...

[www.cs.aau.dk/~dingming](#) - [Cached](#)

[BEWEB 2011 \(BEWEBworkshop\) on Twitter](#)

BEWEB 2011 will be held on March 25 in conjunction with EDBT/ICDT 2011 conf. in Uppsala (Sweden) [http://bit.ly/9U188T](#) #edbt2011 #beweb2011 6:22 AM Nov 15th via web

[twitter.com/BEWEBworkshop](#) - [Cached](#)

[EDBT 2011 : 14th International Conference on Extending ...](#)

[https://cmt.research.microsoft.com/EDBT2011](#) Publication and IP issues Papers will be made available in online conference proceedings, possibly hosted within the ACM ...

[www.wikicfp.com/cfp/servlet/event.showcfp?eventid=10620&...](#) - [Cached](#)

Search Engine Results Page



- ▶ Very important to estimate relevance of the result for the user's query
 - high-quality, query-biased
 - expensive to generate

Different queries....

Web Immagini Video Pagine Gialle Shopping Notizie Altro ▾

university of Uppsala

Cerca: ☒ nel Web ☐ nei siti in italiano

[Uppsala University Sweden](#) - Traduci
Uppsala University is one of northern Europe's most highly ranked universities. ...
Uppsala University pursues research in a wide range of fields that helps us ...
www.uu.se/en - Cache

[Research - Uppsala universitet](#) - Traduci
Uppsala University pursues research in a wide range of fields that helps us ...
Hallberg, Vice Chancellor of Uppsala University. Breadth, diversity, and ...
www.uu.se/en/node4 - Cache

[Uppsala University - Wikipedia, the free encyclopedia](#) - Traduci
[History](#) | [Administration and organisation](#) | [Campus](#) | [Student life](#)
The university rose to pronounced significance during the rise of Sweden as a great power at the end of the 16th century and was then given a relative financial stability with the large...
en.wikipedia.org/wiki/Uppsala_University - 159k - Cache

[University of Uppsala: Information from Answers.com](#) - Traduci
Uppsala, University of, at Uppsala, Sweden; founded 1477 by Sten Sture, the Elder, and Archbishop Jakob Ulvsson ... University of Uppsala. Home > Library > ...
www.answers.com/topic/uppsala-university - 222k - Cache

[Uppsala universitetsbibliotek](#) - Traduci
Uppsala University Library Catalogue " New library catalogue " ... Test: Manual of European Environmental Policy " More news items... Uppsala University Library ...
www.ub.uu.se/eindex.cfm - 119k - Cache

[Uppsala University Hospital - Wikipedia, the free encyclopedia](#) - Traduci
[History](#) | [Present facilities](#) | [Organization](#) | [Uppsala Care](#)
Uppsala University Hospital in Uppsala, Sweden is a teaching hospital for the Uppsala University Faculty of Medicine and the Nursing School. Uppsala University Hospital is owned and...
en.wikipedia.org/wiki/Uppsala_University_Hospital - 59k - Cache



Web Immagini Video Pagine Gialle Shopping Notizie Altro ▾

Uppsala university science

Cerca: ☒ nel Web ☐ nei siti in italiano

[Research - Uppsala universitet](#) - Traduci
Uppsala University pursues research in a wide range of fields that helps us ...
Stockholm-Uppsala Life Science. Codex - rules and regulations for research ...
www.uu.se/en/node4 - Cache

[Master Programme in Computer Science 2010/2011 - Uppsala University Sweden](#) - Traduci
I saw that Uppsala was the oldest university in the Nordic countries and had ... (at Uppsala University), students will study Advanced Computer Science Studies in ...
www.uu.se/en/node605?pKod=TDV2M&lasar=10/11 - Cache

[University of Uppsala: Information from Answers.com](#) - Traduci
Uppsala, University of, at Uppsala, Sweden; founded 1477 by Sten Sture, ... Science. Shopping. Sports. Technology. Travel & Places. Q & A. University of Uppsala ...
www.answers.com/topic/uppsala-university - 222k - Cache

[International Science Programs \(ISP\), Uppsala University, Sweden.](#) - Traduci
International Science Programme, Uppsala University ... Swedish collaborator: Dept. of Medical Cell Biology, Uppsala University (Prof. Erik Gylfe) ...
www.sasnet.lu.se/ispuppsala.html - Cache

[Uppsala University - Wikipedia, the free encyclopedia](#) - Traduci
[History](#) | [Administration and organisation](#) | [Campus](#) | [Student life](#)
The university rose to pronounced significance during the rise of Sweden as a great power at the end of the 16th century and was then given a relative financial stability with the large...
en.wikipedia.org/wiki/Uppsala_University - 159k - Cache

[UU/IT/Research in Computing Science](#) - Traduci
Publications and reports from the Computing Science Division. ... Copyright © 2011 Uppsala University, Department of Information Technology. ...
www.csd.uu.se - Cache



Same title, same URL, same snippet (A)

Web Immagini Video Pagine Gialle Shopping Notizie Altro ▾

university of Uppsala

Cerca: ☒ nel Web ☐ nei siti in italiano

[Uppsala University Sweden](#) - Traduci
Uppsala University is one of northern Europe's most highly ranked universities. ...
Uppsala University pursues research in a wide range of fields that helps us ...
[www.uu.se/en](#) - Cache

[Research - Uppsala universitet](#) - Traduci
Uppsala University pursues research in a wide range of fields that helps us ...
Hallberg, Vice Chancellor of Uppsala University. Breadth, diversity, and ...
[www.uu.se/en/node4](#) - Cache

[Uppsala University - Wikipedia, the free encyclopedia](#) - Traduci
[History](#) | [Administration and organisation](#) | [Campus](#) | [Student life](#)
The university rose to pronounced significance during the rise of Sweden as a great power at the end of the 16th century and was then given a relative financial stability with the large...
[en.wikipedia.org/wiki/Uppsala_University](#) - 159k - Cache

[University of Uppsala: Information from Answers.com](#) - Traduci
Uppsala, University of, at Uppsala, Sweden; founded 1477 by Sten Sture, the Elder, and Archbishop Jakob Ulvsson ... University of Uppsala. Home > Library > ...
[www.answers.com/topic/uppsala-university](#) - 222k - Cache

[Uppsala universitetsbibliotek](#) - Traduci
Uppsala University Library Catalogue " New library catalogue " ... Test: Manual of European Environmental Policy " More news items... Uppsala University Library ...
[www.ub.uu.se/eindex.cfm](#) - 119k - Cache

[Uppsala University Hospital - Wikipedia, the free encyclopedia](#) - Traduci
[History](#) | [Present facilities](#) | [Organization](#) | [Uppsala Care](#)
Uppsala University Hospital in Uppsala, Sweden is a teaching hospital for the Uppsala University Faculty of Medicine and the Nursing School. Uppsala University Hospital is owned and...
[en.wikipedia.org/wiki/Uppsala_University_Hospital](#) - 59k - Cache



Web Immagini Video Pagine Gialle Shopping Notizie Altro ▾

Uppsala university science

Cerca: ☒ nel Web ☐ nei siti in italiano

[Research - Uppsala universitet](#) - Traduci
Uppsala University pursues research in a wide range of fields that helps us ...
Stockholm-Uppsala Life Science. Codex - rules and regulations for research ...
[www.uu.se/en/node4](#) - Cache

[Master Programme in Computer Science 2010/2011 - Uppsala University Sweden](#) - Traduci
I saw that Uppsala was the oldest university in the Nordic countries and had ... (at Uppsala University), students will study Advanced Computer Science Studies in ...
[www.uu.se/en/node605?pKod=TDV2M&lasar=10/11](#) - Cache

[University of Uppsala: Information from Answers.com](#) - Traduci
Uppsala, University of, at Uppsala, Sweden; founded 1477 by Sten Sture, ... Science. Shopping. Sports. Technology. Travel & Places. Q & A. University of Uppsala ...
[www.answers.com/topic/uppsala-university](#) - 222k - Cache

[International Science Programs \(ISP\), Uppsala University, Sweden.](#) - Traduci
International Science Programme, Uppsala University ... Swedish collaborator: Dept. of Medical Cell Biology, Uppsala University (Prof. Erik Gylfe) ...
[www.sasnet.lu.se/ispuppsala.html](#) - Cache

[Uppsala University - Wikipedia, the free encyclopedia](#) - Traduci
[History](#) | [Administration and organisation](#) | [Campus](#) | [Student life](#)
The university rose to pronounced significance during the rise of Sweden as a great power at the end of the 16th century and was then given a relative financial stability with the large...
[en.wikipedia.org/wiki/Uppsala_University](#) - 159k - Cache

[UU/IT/Research in Computing Science](#) - Traduci
Publications and reports from the Computing Science Division. ... Copyright © 2011 Uppsala University, Department of Information Technology. ...
[www.csd.uu.se](#) - Cache



Same title, same URL, different snippets (B)

Web Immagini Video Pagine Gialle Shopping Notizie Altro ▾

university of Uppsala

Cerca: ☒ nel Web ☐ nei siti in italiano

[Uppsala University Sweden](#) - Traduci
Uppsala University is one of northern Europe's most highly ranked universities. ...
Uppsala University pursues research in a wide range of fields that helps us ...
www.uu.se/en - Cache

[Research - Uppsala universitet](#) - Traduci
Uppsala University pursues research in a wide range of fields that helps us ...
Hallberg, Vice Chancellor of Uppsala University. Breadth, diversity, and ...
www.uu.se/en/node4 - Cache

[Uppsala University - Wikipedia, the free encyclopedia](#) - Traduci
[History](#) | [Administration and organisation](#) | [Campus](#) | [Student life](#)
The university rose to pronounced significance during the rise of Sweden as a great power at the end of the 16th century and was then given a relative financial stability with the large...
en.wikipedia.org/wiki/Uppsala_University - 159k - Cache

[University of Uppsala: Information from Answers.com](#) - Traduci
Uppsala, University of, at Uppsala, Sweden; founded 1477 by Sten Sture, the Elder, and Archbishop Jakob Ulvsson ... University of Uppsala. Home > Library > ...
www.answers.com/topic/uppsala-university - 222k - Cache

[Uppsala universitetsbibliotek](#) - Traduci
Uppsala University Library Catalogue " New library catalogue " ... Test: Manual of European Environmental Policy " More news items... Uppsala University Library ...
www.ub.uu.se/eindex.cfm - 119k - Cache

[Uppsala University Hospital - Wikipedia, the free encyclopedia](#) - Traduci
[History](#) | [Present facilities](#) | [Organization](#) | [Uppsala Care](#)
Uppsala University Hospital in Uppsala, Sweden is a teaching hospital for the Uppsala University Faculty of Medicine and the Nursing School. Uppsala University Hospital is owned and...
en.wikipedia.org/wiki/Uppsala_University_Hospital - 59k - Cache



Web Immagini Video Pagine Gialle Shopping Notizie Altro ▾

Uppsala university science

Cerca: ☒ nel Web ☐ nei siti in italiano

[Research - Uppsala universitet](#) - Traduci
Uppsala University pursues research in a wide range of fields that helps us ...
Stockholm-Uppsala Life Science. Codex - rules and regulations for research ...
www.uu.se/en/node4 - Cache

[Master Programme in Computer Science 2010/2011 - Uppsala University Sweden](#) - Traduci
I saw that Uppsala was the oldest university in the Nordic countries and had ... (at Uppsala University), students will study Advanced Computer Science Studies in ...
www.uu.se/en/node605?pKod=TDV2M&lasar=10/11 - Cache

[University of Uppsala: Information from Answers.com](#) - Traduci
Uppsala, University of, at Uppsala, Sweden; founded 1477 by Sten Sture, ... Science. Shopping. Sports. Technology. Travel & Places. Q & A. University of Uppsala ...
www.answers.com/topic/uppsala-university - 222k - Cache

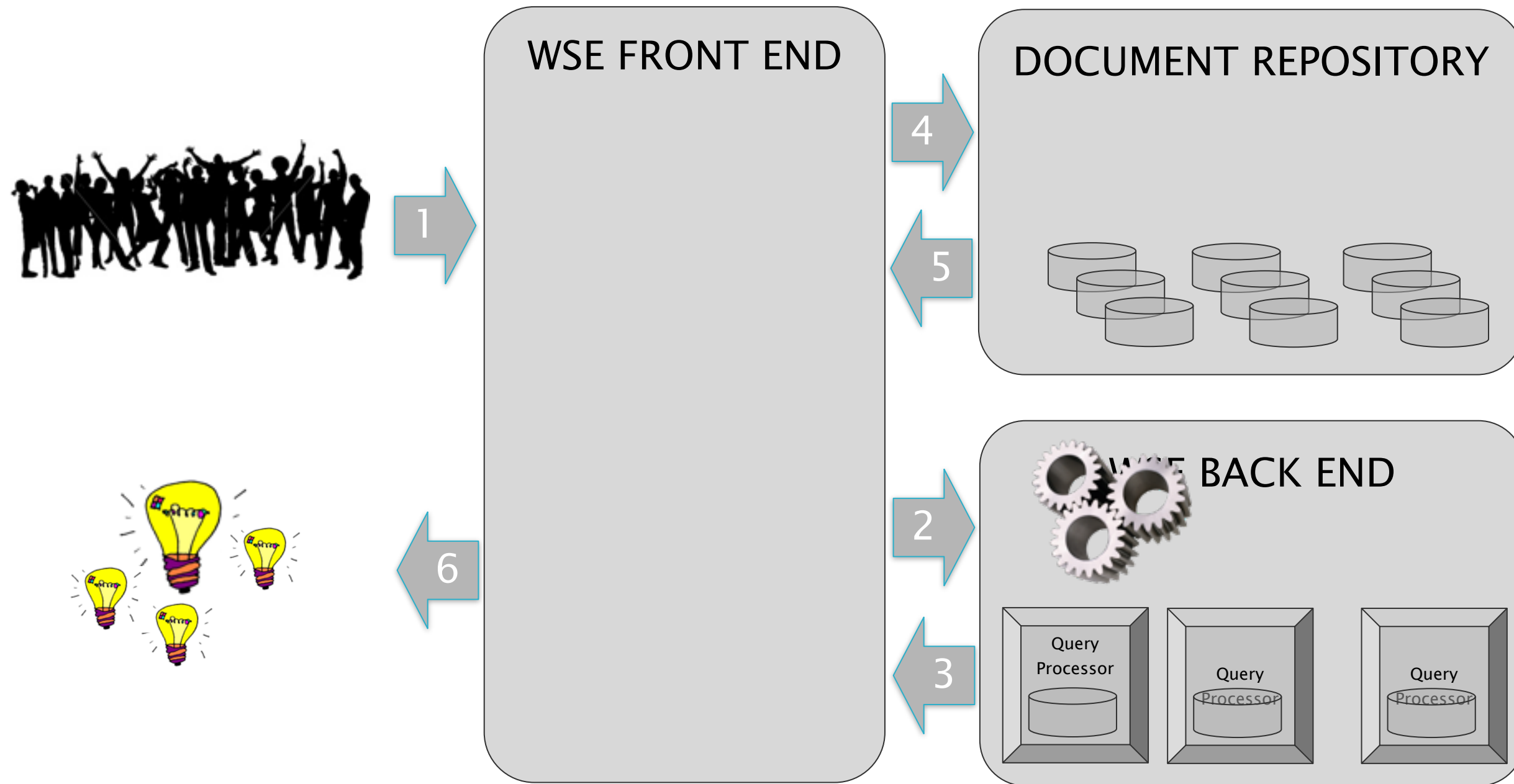
[International Science Programs \(ISP\), Uppsala University, Sweden.](#) - Traduci
International Science Programme, Uppsala University ... Swedish collaborator: Dept. of Medical Cell Biology, Uppsala University (Prof. Erik Gylfe) ...
www.sasnet.lu.se/ispuppsala.html - Cache

[Uppsala University - Wikipedia, the free encyclopedia](#) - Traduci
[History](#) | [Administration and organisation](#) | [Campus](#) | [Student life](#)
The university rose to pronounced significance during the rise of Sweden as a great power at the end of the 16th century and was then given a relative financial stability with the large...
en.wikipedia.org/wiki/Uppsala_University - 159k - Cache

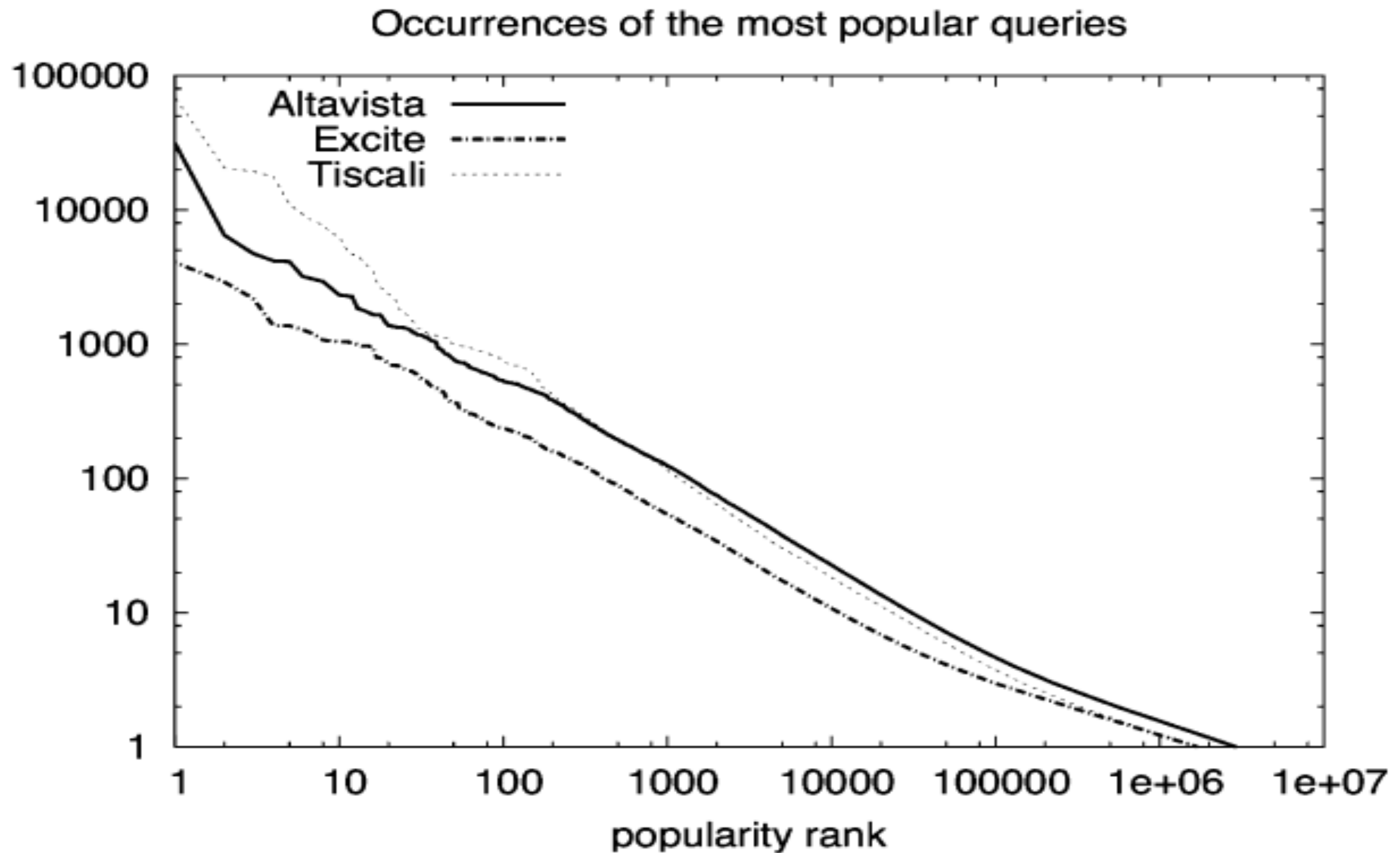
[UU/IT/Research in Computing Science](#) - Traduci
Publications and reports from the Computing Science Division. ... Copyright © 2011 Uppsala University, Department of Information Technology. ...
www.csd.uu.se - Cache



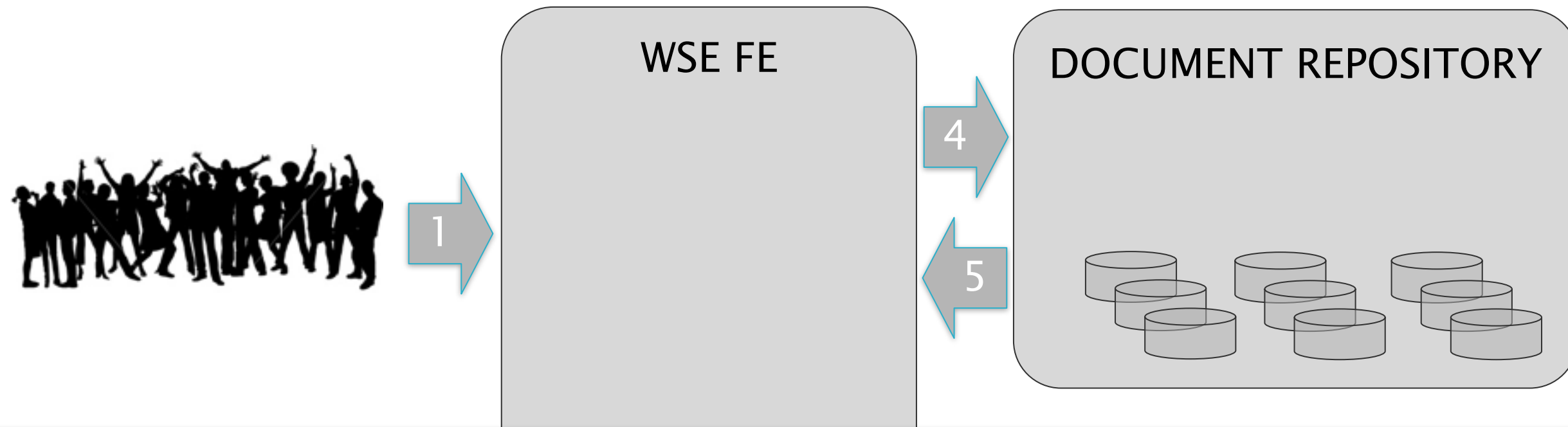
Web Searching process



WSE Front End caching

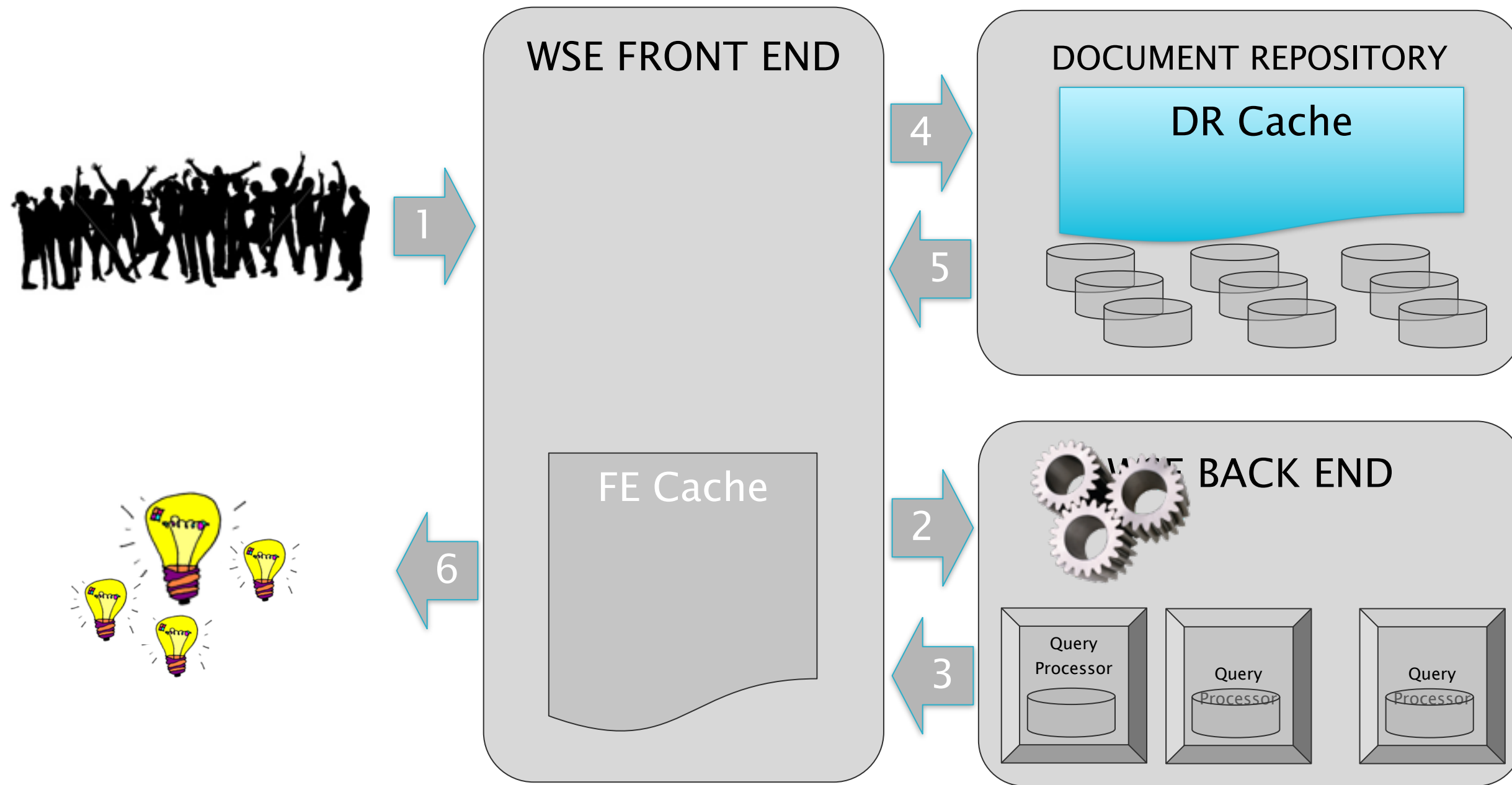


WSE Front End caching



SERPs' caching does not help in the cases A and B above, when the same or a similar snippet is generated for the same doc, but for slightly different queries

Doc Repository Caching strategies



Doc Repository Caching

Given q and $docID$, the DR has to: 1) retrieve **document content** and **URL**, 2) generate an effective query-biased snippet

- ▶ A DR cache can substantially reduce disk accesses
- ▶ Two cache organizations discussed in literature:
 - $DR_{Cachedoc}$ storing integral copies of most accessed documents*
 - $DR_{Cachesurr}$ storing surrogates of most accessed documents**

We propose $DR_{CacheSsnip}$ whose entries store **supersnippets** built on the basis of the past queries submitted to the WSE

*Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E. Williams. Fast generation of result snippets in web search. SIGIR '07

**Y. Tsegay, S. J. Puglisi, A. Turpin, and J. Zobel, Document Compaction for Efficient Query Biased Snippet Generation. ECIR '09

The Query Log used

► MSN RFP 2006 query log

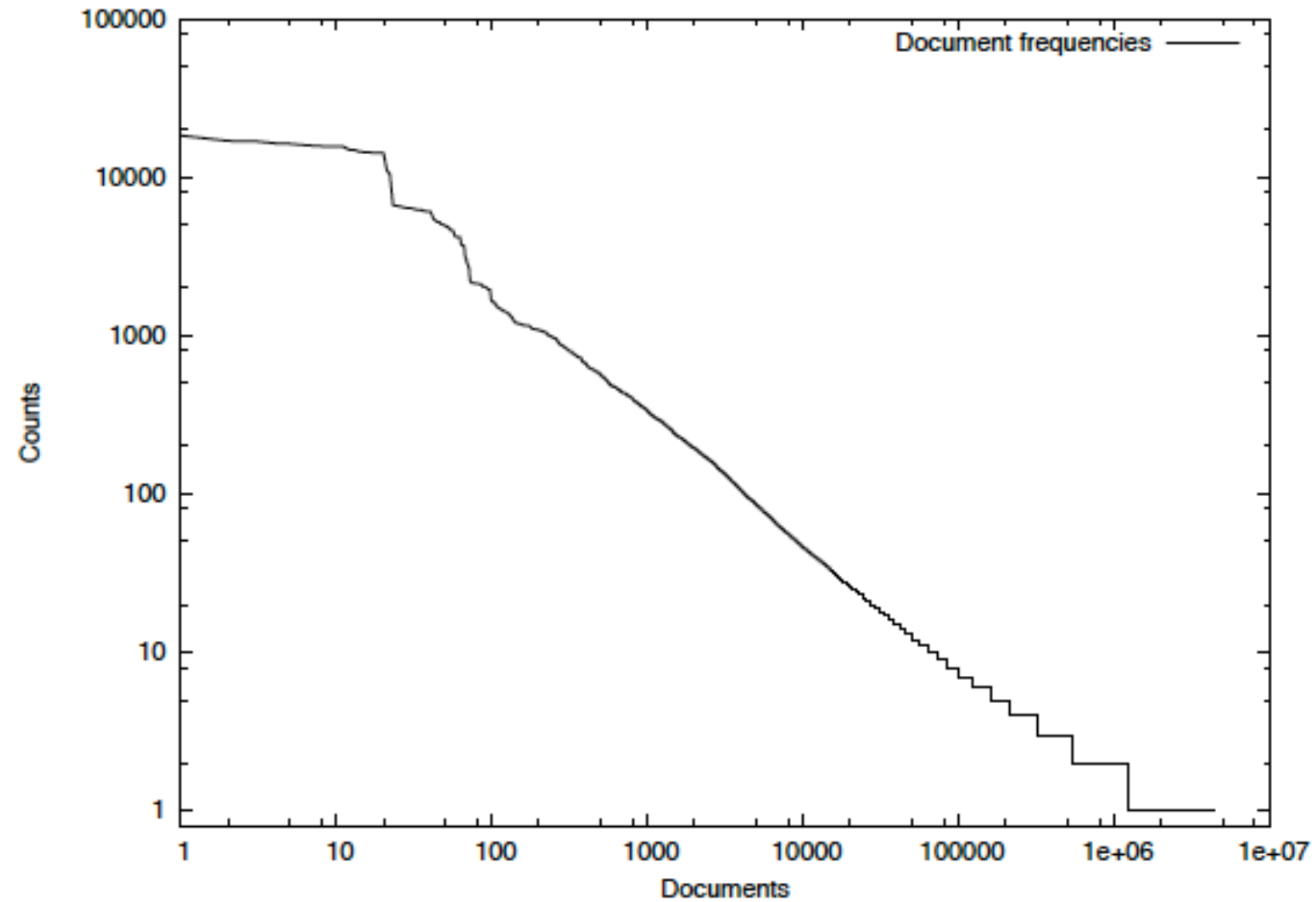
- queries from a US Microsoft search site sampled over one month (May 2006)
- submitted to Yahoo! BOSS to retrieve top-10 results

- 9,000,000 queries
- 4,447,444 distinct queries
- 25,897,247 distinct URLs → high URLs sharing!

Query Log	Queries	Distinct Queries	Distinct Urls (top 10)
D1			
<i>training</i>	1,000,000	601,369	4,317,603
<i>test</i>	500,000	310,495	2,324,295
D2			
<i>training</i>	9,000,000	4,447,444	25,897,247
<i>test</i>	4,500,000	2,373,227	12,134,453

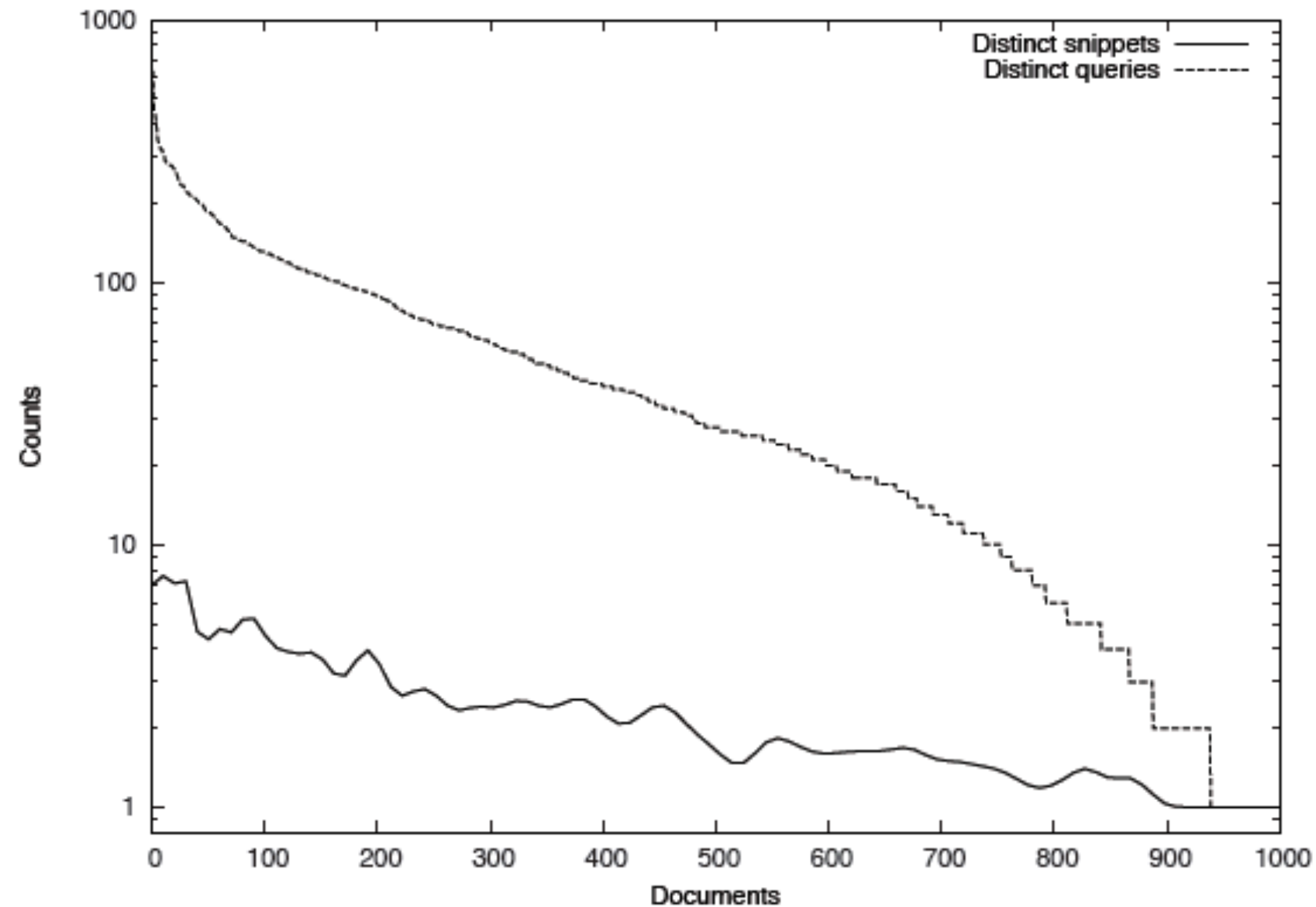
Fact 1: URLs occurrences in SERPs

- ▶ Double Pareto distribution



Fact 2: top-1000 most frequently retrieved documents

- ▶ snippets generated for the same document are a few (<9)
- ▶ a document may answer several different queries



Evidences

1. Relevant documents are characterized by a few snippets
2. Relevant documents are shared by different queries
3. Snippets are made of a few sentences
4. Different snippets from the same document share common sentences

Q: What do you get when you cross
an Elephant with a Rhino?



A: I haven't a goddamn clue.

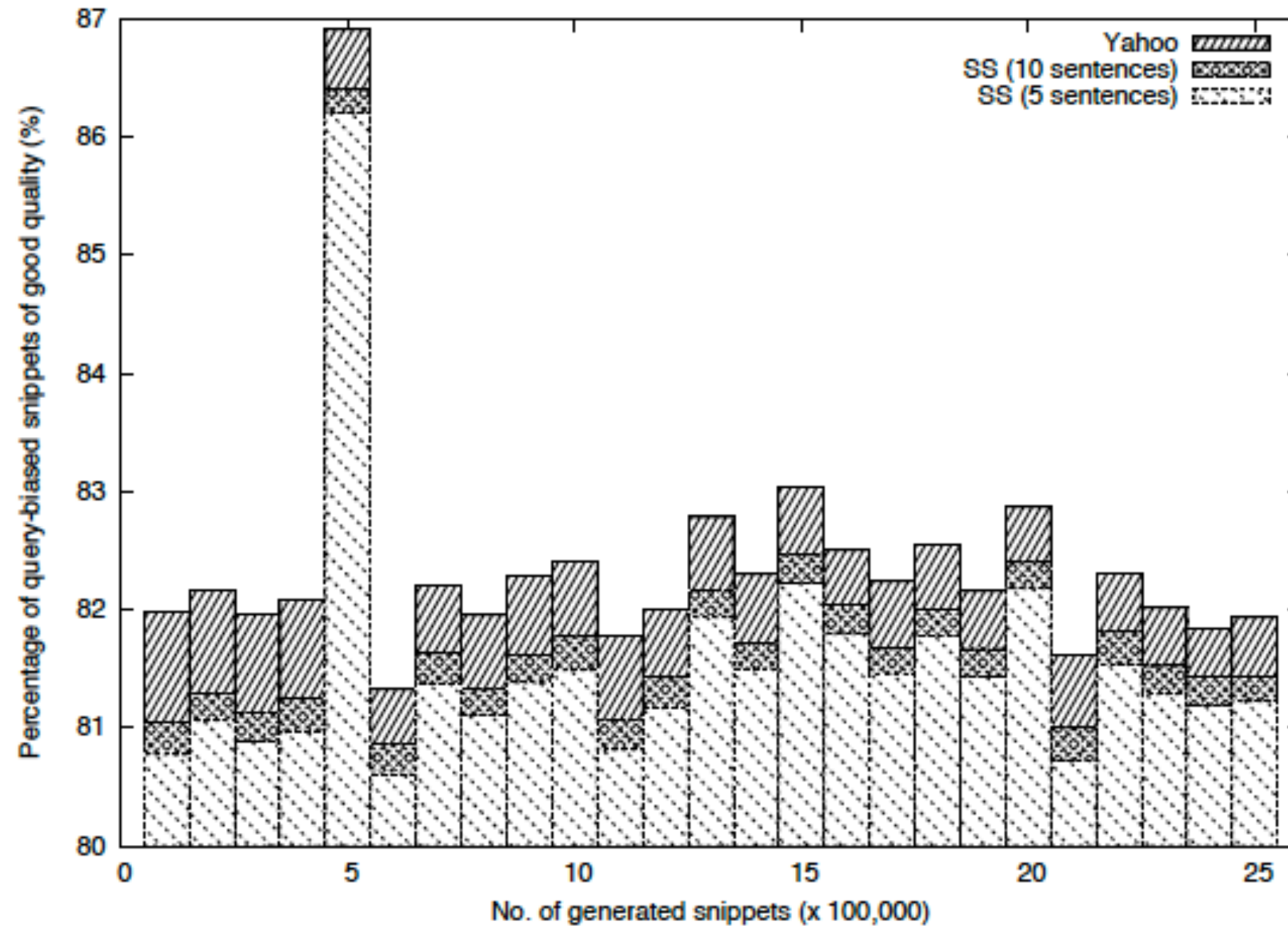
Definition of QL-based supersnippet

Given the set Q_d of all the past queries in QL for which document d was returned, we define supersnippet ss_n of d the set of the n most frequent sentences occurring in the snippets $S_{d,q}$ generated for answering the queries in Q_d

Claim

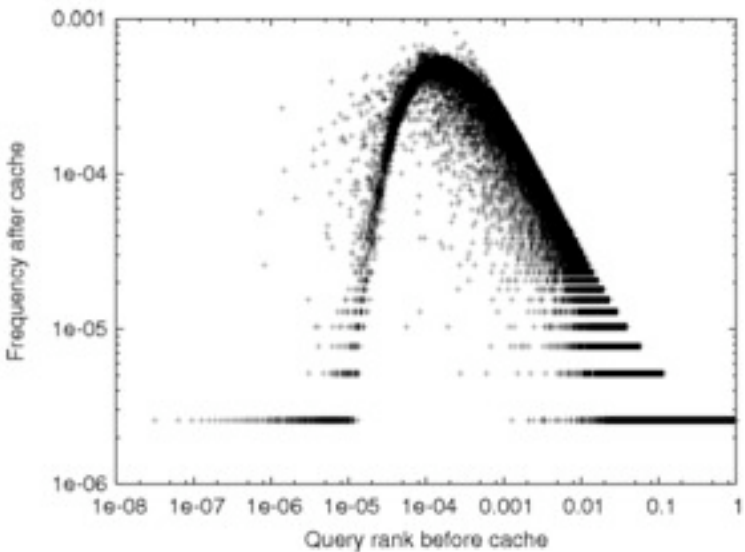
Caching LRU supersnippets ss_n is an effective and efficient strategy for DR_{Cache}

Effectiveness: $DR_{\text{CacheSsnip}}$ vs. Yahoo! BOSS



Efficiency

FECache #Entries (Hit Ratio)	DRCache	Size (in MB)	Hit Ratio DRCache
SERPs cache filtering out most recently submitted queries 256K (0.38)	<i>doc</i>	256M	0.083
		512M	0.1
		1024M	0.134
		2048M	0.17
	<i>surr</i>	256M	0.104
		512M	0.135
		1024M	0.177
		2048M	0.23
	<i>ss5</i>	256M	0.139
		512M	0.194
		1024M	0.266
		2048M	0.34
	<i>ss10</i>	256M	0.13
		512M	0.189
		1024M	0.266
		2048M	0.334



SS Cache peculiarities

CacheLookup(q, DocID)

if InCache(DocID) then

SS $\leftarrow$ RetrieveSS(DocID)

snippet $\leftarrow$ GenerateS(q,SS)

if Quality(q,SS) < Threshold then

 snippet $\leftarrow$ UpdateSS (SS, DocID, q)

UpdateLRU(DocID)

else

SS $\leftarrow$ null

snippet $\leftarrow$ UpdateSS (SS, DocID, q)

ReplaceLRU(SS, DocID)

SS Cache peculiarities

CacheLookup(q, DocID)

if InCache(DocID) then

SS $\leftarrow$ RetrieveSS(DocID)

snippet $\leftarrow$ GenerateS(q,SS)

if Quality(q,SS) < Threshold then

snippet $\leftarrow$ UpdateSS (SS, DocID, q)

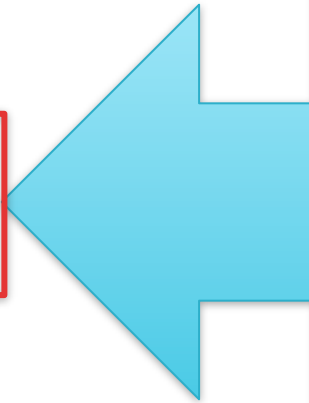
UpdateLRU(DocID)

else

SS $\leftarrow$ null

snippet $\leftarrow$ UpdateSS (SS, DocID, q)

ReplaceLRU(SS, DocID)



A “Quality” Miss
may occur when
the quality of the
snippet generated
from the available
SS is below a
given threshold

Quality Miss count

- ▶ We count a Quality Miss when even a single query term is not present in the snippet
 - Also Yahoo! BOSS snippets sometimes do not contain ALL query terms
 - misspellings, disjunctive queries, etc.
- ▶ We underestimate the Hit Ratio

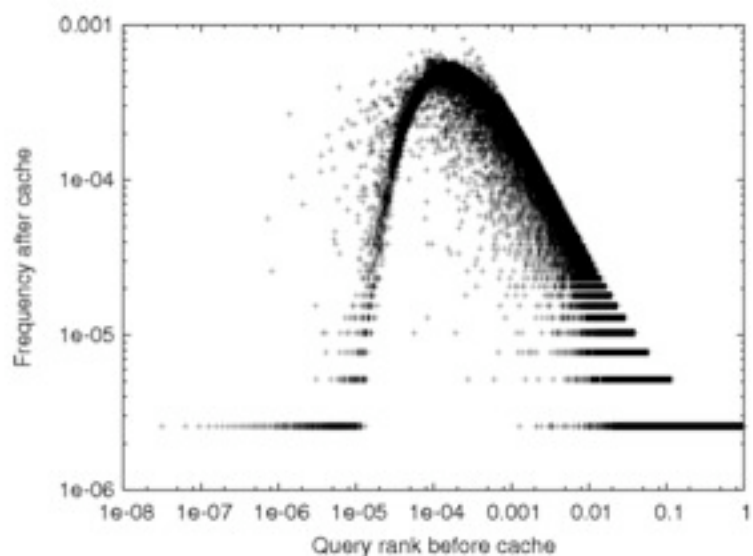
What about considering a Quality Miss only when our snippets' quality is lower than Yahoo!'s one ?

Refined Efficiency Measurement

SERPs cache
filtering out most
recently submitted
queries



128K (0.36)



DRCache	Size (in MB)	Hit Ratio DRCache
SS5	256M	0.21
	512M	0.29
	1024M	0.36
	2048M	0.42
SS10	256M	0.2
	512M	0.28
	1024M	0.356
	2048M	0.419
	256M	0.19

Total Hit Ratio: ~62%

Conclusion & Future Work

- ▶ Novel technique for scaling up search engine performance by snippet caching
 - motivated by the analysis of a large real-world query log
 - enabling the generation of effective snippets even for queries and URLs not previously seen
 - achieving high Hit Ratios (cumulatively up to 0.62!)
 - evaluated by means of large-scale experiments
- ▶ Open issues
 - study of dynamic vs static policies for managing supersnippets
 - quality of SSs suffers for aging?
 - study of diversification in supersnippets?
 - heuristics, machine learning approaches for managing Quality Misses

Questions?

