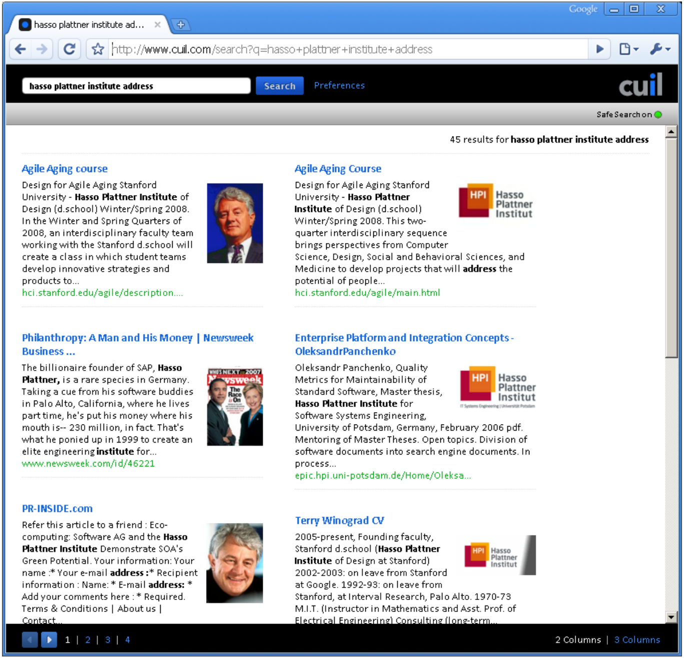
Fast Generation of Result Snippets in Web Search

Felix Geller

January 20, 2009

Agenda

- What are “snippets”?
- How do they fit in?
- Possible difficulties
- Speeding up snippet generation
  - Document compression
  - Document compaction
- Summary


hasso plattner institute ad...
http://www.cuil.com/search?q=hasso+plattner+institute+address
hasso plattner institute address
Search
Preferences
cuil
Safe Search on

45 results for **hasso plattner institute address**

Agile Aging course

Design for Agile Aging Stanford University - **Hasso Plattner Institute** of Design (d.school) Winter/Spring 2008. In the Winter and Spring Quarters of 2008, an interdisciplinary faculty team working with the Stanford d.school will create a class in which student teams develop innovative strategies and products to...
hci.stanford.edu/agile/description....

Agile Aging Course

Design for Agile Aging Stanford University - **Hasso Plattner Institute** of Design (d.school) Winter/Spring 2008. This two-quarter interdisciplinary sequence brings perspectives from Computer Science, Design, Social and Behavioral Sciences, and Medicine to develop projects that will **address** the potential of people...
hci.stanford.edu/agile/main.html

Philanthropy: A Man and His Money | Newsweek Business ...

The billionaire founder of SAP, **Hasso Plattner,** is a rare species in Germany. Taking a cue from his software buddies in Palo Alto, California, where he lives part time, he's put his money where his mouth is-- 230 million, in fact. That's what he ponied up in 1999 to create an elite engineering **institute** for...
www.newsweek.com/id/46221

Enterprise Platform and Integration Concepts - OleksandrPanchenko

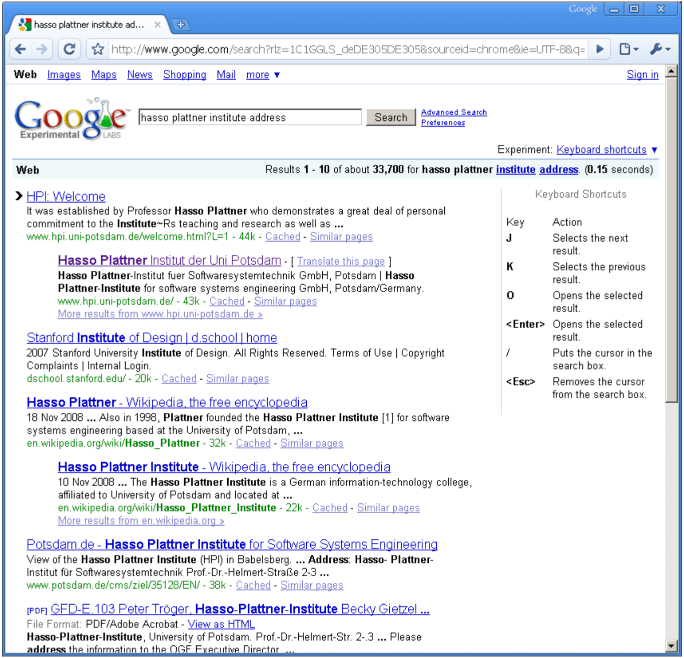
Oleksandr Panchenko, Quality Metrics for Maintainability of Standard Software, Master thesis, **Hasso Plattner Institute** for Software Systems Engineering, University of Potsdam, Germany, February 2006 pdf. Mentoring of Master Theses. Open topics. Division of software documents into search engine documents. In process...
epic.hpi.uni-potsdam.de/Home/Oleksa...

PR-INSIDE.com

Refer this article to a friend : Eco-computing: Software AG and the **Hasso Plattner Institute** Demonstrate SOA's Green Potential. Your information: Your name :* Your e-mail **address :*** Recipient information : Name: * E-mail **address:** * Add your comments here : * Required. Terms & Conditions | About us | Contact...

Terry Winograd CV

2005-present, Founding faculty, Stanford d.school (**Hasso Plattner Institute** of Design at Stanford) 2002-2003: on leave from Stanford at Google. 1992-93: on leave from Stanford, at Interval Research, Palo Alto. 1970-73 M.I.T. (Instructor in Mathematics and Asst. Prof. of Electrical Engineering) Consulting (long-term...


hasso plattner institute ad...
http://www.google.com/search?rlz=1C1GGLS_deDE305DE305&sourceid=chrome&ie=UTF-8&q=
Web Images Maps News Shopping Mail more ▾
Sign in
Google Experimental LABS
hasso plattner institute address
Search
Advanced Search
Preferences
Experiment: Keyboard shortcuts ▾

Web Results **1** - **10** of about **33,700** for **hasso plattner institute address**. (**0.15** seconds)

> HPI: Welcome
It was established by Professor **Hasso Plattner** who demonstrates a great deal of personal commitment to the **Institute**~Rs teaching and research as well as **...**
www.hpi.uni-potsdam.de/welcome.html?L=1 - 44k - Cached - Similar pages

Hasso Plattner Institut der Uni Potsdam - [Translate this page]
Hasso Plattner-Institut fuer Softwaresystemtechnik GmbH, Potsdam | **Hasso Plattner-Institute** for software systems engineering GmbH, Potsdam/Germany.
www.hpi.uni-potsdam.de/ - 43k - Cached - Similar pages
More results from www.hpi.uni-potsdam.de »

Stanford **Institute** of Design | d.school | home
2007 Stanford University **Institute** of Design. All Rights Reserved. Terms of Use | Copyright Complaints | Internal Login.
dschool.stanford.edu/ - 20k - Cached - Similar pages

Hasso Plattner - Wikipedia, the free encyclopedia
18 Nov 2008 **...** Also in 1998, **Plattner** founded the **Hasso Plattner Institute** [1] for software systems engineering based at the University of Potsdam, **...**
en.wikipedia.org/wiki/**Hasso_Plattner** - 32k - Cached - Similar pages

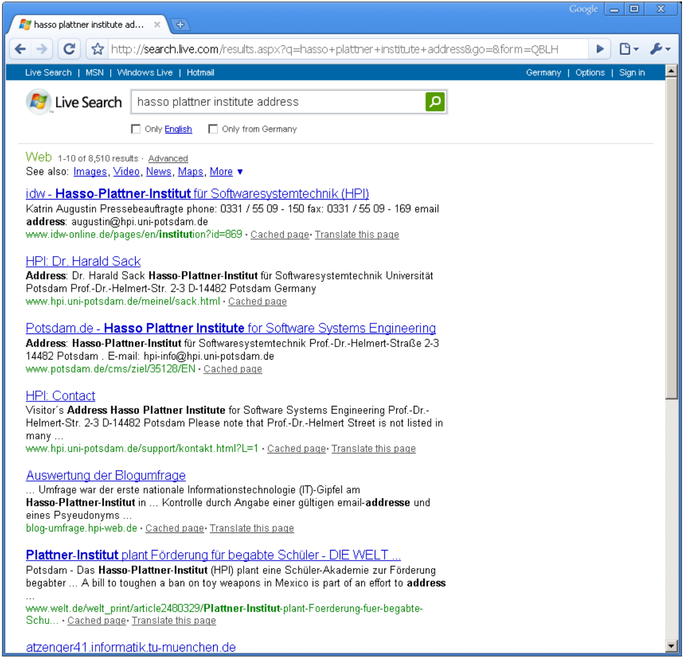
Hasso Plattner Institute - Wikipedia, the free encyclopedia
10 Nov 2008 **...** The **Hasso Plattner Institute** is a German information-technology college, affiliated to University of Potsdam and located at **...**
en.wikipedia.org/wiki/**Hasso_Plattner_Institute** - 22k - Cached - Similar pages
More results from en.wikipedia.org »

Potsdam.de - **Hasso Plattner Institute** for Software Systems Engineering
View of the **Hasso Plattner Institute** (HPI) in Babelsberg. **... Address**: **Hasso- Plattner-** Institut für Softwaresystemtechnik Prof.-Dr.-Helmert-Straße 2-3 **...**
www.potsdam.de/cms/ziel/35128/EN/ - 38k - Cached - Similar pages

[PDF] GFD-E.103 Peter Tröger, **Hasso-Plattner-Institute** Becky Gietzel **...**
File Format: PDF/Adobe Acrobat - View as HTML
Hasso-Plattner-Institute, University of Potsdam. Prof.-Dr.-Helmert-Str. 2-.3 **...** Please **address** the information to the OGF Executive Director **...**

Keyboard Shortcuts

Key	Action
J	Selects the next result.
K	Selects the previous result.
O	Opens the selected result.
<Enter>	Opens the selected result.
/	Puts the cursor in the search box.
<Esc>	Removes the cursor from the search box.


hasso plattner institute ad...
Google
http://search.live.com/results.aspx?q=hasso+plattner+institute+address&go=&form=QBLH
Live Search | MSN | Windows Live | Hotmail
Germany | Options | Sign in
Live Search
hasso plattner institute address
□ Only English □ Only from Germany

Web 1-10 of 8,510 results · Advanced

See also: Images, Video, News, Maps, More ▾

idw - **Hasso-Plattner-Institut** für Softwaresystemtechnik (HPI)
Katrin Augustin Pressebeauftragte phone: 0331 / 55 09 - 150 fax: 0331 / 55 09 - 169 email **address**: augustin@hpi.uni-potsdam.de
www.idw-online.de/pages/en/**institut**ion?id=869 · Cached page· Translate this page

HPI: Dr. Harald Sack
Address: Dr. Harald Sack **Hasso-Plattner-Institut** für Softwaresystemtechnik Universität Potsdam Prof.-Dr.-Helmert-Str. 2-3 D-14482 Potsdam Germany
www.hpi.uni-potsdam.de/meinel/sack.html · Cached page

Potsdam.de - **Hasso Plattner Institute** for Software Systems Engineering
Address: **Hasso-Plattner-Institut** für Softwaresystemtechnik Prof.-Dr.-Helmert-Straße 2-3 14482 Potsdam . E-mail: hpi-info@hpi.uni-potsdam.de
www.potsdam.de/cms/ziel/35128/EN · Cached page

HPI: Contact
Visitor´s **Address Hasso Plattner Institute** for Software Systems Engineering Prof.-Dr.-Helmert-Str. 2-3 D-14482 Potsdam Please note that Prof.-Dr.-Helmert Street is not listed in many ...
www.hpi.uni-potsdam.de/support/kontakt.html?L=1 · Cached page· Translate this page

Auswertung der Blogumfrage
... Umfrage war der erste nationale Informationstechnologie (IT)-Gipfel am **Hasso-Plattner-Institut** in ... Kontrolle durch Angabe einer gültigen email-**addresse** und eines Psyeudonyms ...
blog-umfrage.hpi-web.de · Cached page· Translate this page

Plattner-Institut plant Förderung für begabte Schüler - DIE WELT ...
Potsdam - Das **Hasso-Plattner-Institut** (HPI) plant eine Schüler-Akademie zur Förderung begabter ... A bill to toughen a ban on toy weapons in Mexico is part of an effort to **address** ...
www.welt.de/welt_print/article2480329/**Plattner-Institut**-plant-Foerderung-fuer-begabte-Schu... · Cached page· Translate this page

atzenger41.informatik.tu-muenchen.de

Query-Biased Snippets

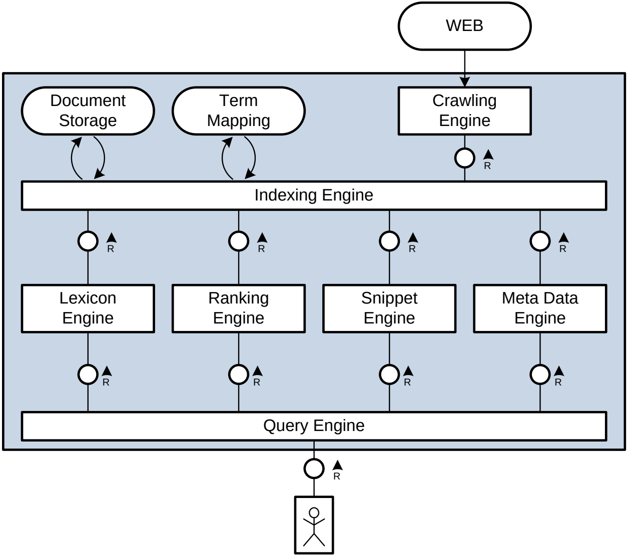
> **Snippets** *are short fragments of text extracted from the document content (or its metadata). ... A* **query-biased** *snippet is one selectively extracted on the basis of its relation to the searcher's query.* **[4]**

→ **Quickly** identify relevant documents **without opening** the document as a whole.

Agenda

- What are “snippets”?
- **How do they fit in?**
- Possible difficulties
- Speeding up snippet generation
  - Document compression
  - Document compaction
- Summary

Abstract Search Engine Architecture



Agenda

- What are “snippets”?
- How do they fit in?
- **Possible difficulties**
- Speeding up snippet generation
  - Document compression
  - Document compaction
- Summary

Possible Difficulties

Relevance Query-biased, i.e. non-static summary.

Context "John McCarthy Ph.D. 1951 –
Creator of the **LISP** Programming Language." [5]

Speed
- Storage: "order of ten billion web pages" [4]
- Load: "hundreds of millions of search queries per day" [4]
- Response: File I/O is a major bottleneck

Agenda

- What are “snippets”?
- How do they fit in?
- Possible difficulties
- **Speeding up snippet generation**
  - Document compression
  - Document compaction
- Summary

Speed → Use Caches

[M]ajority of time spent generating a snippet is in locating the document on disk . . . : 64% for whole documents.

*With 1% of documents cached, . . . around 80% of disk seeks are avoided.***[4]**

Disk Cache	Managed by OS, e.g. stores frequently accessed documents
Query Cache	Stores precomputed result pages for popular queries
Document Cache	Stores frequently accessed documents in main memory

Agenda

- What are “snippets”?
- How do they fit in?
- Possible difficulties
- Speeding up snippet generation
  - **Document compression**
  - Document compaction
- Summary

Document Compression: Concepts

- Compressed Token System (CTS)
- Document content is normalized (convert `br`, remove tags)
- Atomic entity: Word
- Entity of interest: Sentence
- Replace words with numbers
- vbyte coding scheme (think UTF-8)
- Words alternate non-words (i.e. punctuation)

Document Compression: Algorithm

1st Pass

- Collect words
- Collect non-words
- Construct model

2nd Pass

- Replace words and non-words
- Escape words which are not encoded

Document Compression: Example

Educators, generals, dieticians, psychologists, and parents program.

Armies, students, and some societies are programmed.

An assault on large problems employs a succession of programs, most of which spring into existence en route.

These programs are rife with issues that appear to be particular to the problem at hand.

To appreciate programming as an intellectual activity in its own right you must turn to computer programming;

you must read and write computer programs -- many of them.

It doesn’t matter much what the programs are about or what applications they serve.

What does matter is how well they perform and how smoothly they fit with other programs in the creation of still greater programs.

Sample taken from [1]

Document Compression: Example

Word Model

Code	Word
0	"with"
1	"you"
2	"how"
3	"are"
4	"in"
5	"computer"
...	...

Non-Word Model

Code	Non-word
0	" "
1	"."
2	", "
3	" _ "
4	"""
5	";"

Document Compression: Example

```
|Educators2|generals2|dieticians2|psychologists260|parents0|program1

|Armies2|students260|some0|societies030|programmed1

|An0|assault0|on0|large0|problems0|employs0|a0|succession0140112|most
     0140|which0|spring0|into0|existence0|en0|route1

|These011030|rife000|issues0|that0|appear090|be0|particular090120
     |problem0|at0|hand1

|To0|appreciate0100|as0|an0|intellectual0|activity040|its0|own0|right
     010130|turn09050105

10130|read060|write050113|many0140|them1

|It0|doesn4|t080|much07012011030|about0|or070|applications0150|serve1

|What0|does080|is020|well0150|perform06020|smoothly0150|fit000|other
     011040120|creation0140|still0|greater0111
```

Document Compression: Gain I/II

	WT10G	WT50G	WT100G
No. Docs ($\times 10^6$)	1.7	10.1	18.5
Raw Text	10,522 MB	56,684 MB	102,833 MB
Baseline (*zlib*)	24%	19%	19%
CTS (+1024 MB)	26%	21%	22%

Taken from [4]

Document Compression: Gain II/II

	WT10G	WT50G	WT100G
Baseline	75	157	183
CTS	38	70	77
Reduction in time	49%	56%	58%

Average time (ms) for the final 7000 queries.

Taken from [4]

Agenda

- What are “snippets”?
- How do they fit in?
- Possible difficulties
- Speeding up snippet generation
  - Document compression
  - **Document compaction**
- Summary

Document Compaction: Concepts

- Reduce size of documents
  → Remove sentences which are deemed insignificant
- Reduce query time
  → Order sentences by significance

Document Compaction: Techniques

Natural order First sentence should introduce paragraph!

Significant terms (ST) Score based on term frequency [2]

Query log based (QLt) Score based on past query terms

Query log based (QLu) Same as QLt, but considers only unique terms

Intermezzo: Sentence Ranking

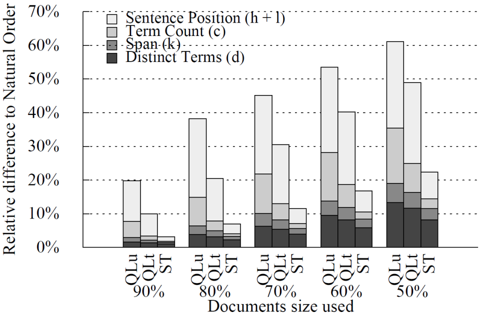
Document is broken into sentences S where $S = [w_1, w_2, ..., w_m]$.
Query Q where $Q = \{q_1, q_2, ..., q_n\}$

h	sentence is a heading
l	sentence is first or second line of document
k	length of longest contiguous run of q_i in S
c	count of $w_j \in Q$
d	c minus repititions

Impact on Sentence Ranking

Impact of omitting 50% of sentences based on ST

h + *l*	approximately 8% change
c	approximately 2% change
k	approximately 3% change
d	approximately 8% change



Taken from [4]

Agenda

- What are “snippets”?
- How do they fit in?
- Possible difficulties
- Speeding up snippet generation
  - Document compression
  - Document compaction
- **Summary**

Summary

- What are “snippets”?
  ⟶ Query-biased text fragments, facilitating the identification of relevant information.
- Possible difficulties
  ⟶ Relevance, Context, Speed.
- Speeding up snippet generation ⟶ Make use of caches.
  - Document compression
    ⟶ Encode words using numbers, make use of dictionary.
  - Document compaction
    ⟶ Reducing document size by 50% has arguably low impact.

References

Harold Abelson, Gerald Sussman, and Julie Sussman.
Structure and Interpretation of Computer Programs.
McGraw-Hill Higher Education, 1996.

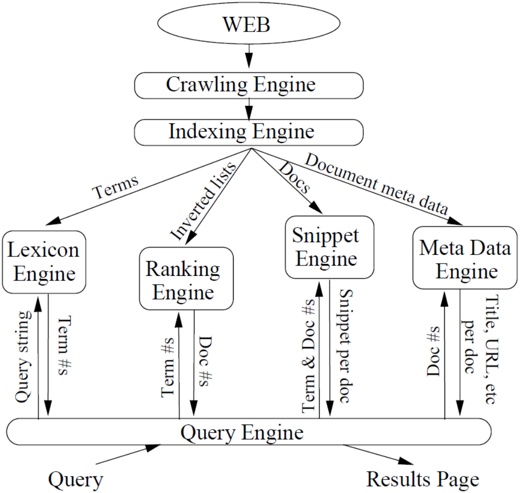
H.P. Luhn.
The automatic creation of literature abstracts.
IBM Journal, 2:159–165, 1958.

Anastasios Tombros and Mark Sanderson.
Advantages of query biased summaries in information retrieval.
In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–10, New York, NY, USA, 1998. ACM.

Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E. Williams.
Fast generation of result snippets in web search.
In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–134, New York, NY, USA, 2007. ACM.

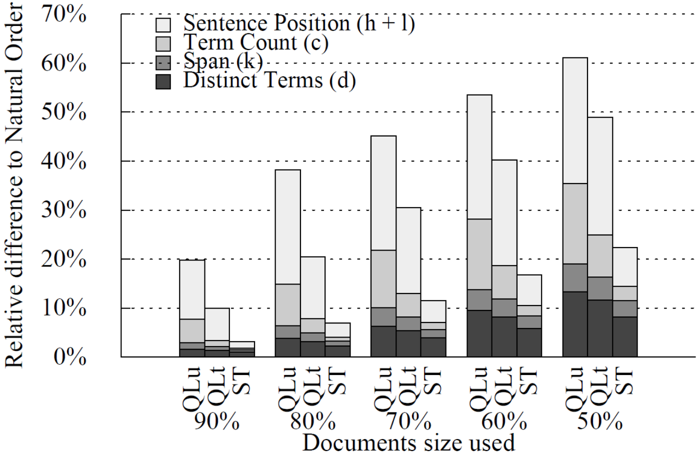
Wikipedia.
List of princeton university people — Wikipedia, the free encyclopedia, 2009.
[Online; accessed 14-January-2009].

Backup Slide



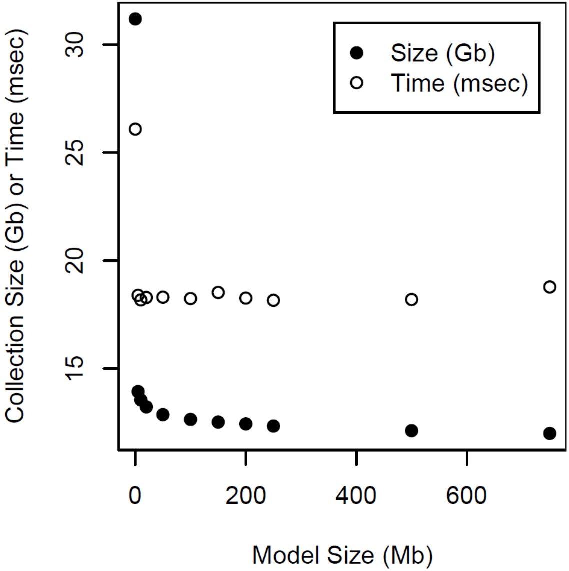
Taken from [4]

Backup Slide



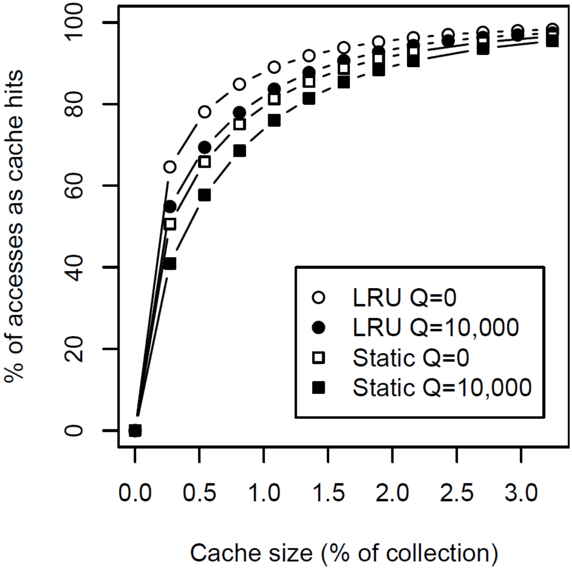
Taken from [4]

Backup Slide



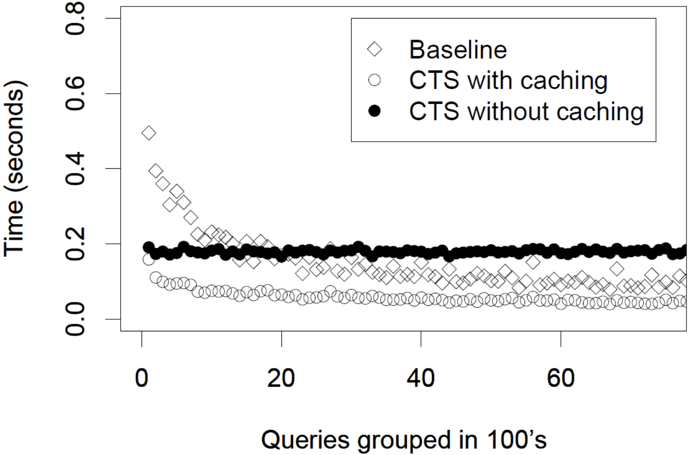
Taken from [4]

Backup Slide



Taken from [4]

Backup Slide



Taken from [4]

Backup Slide

Let $f_{d,t}$ be the frequency of term t in document d, then term t is determined to be significant if:

$$f_{d,t} \leq \begin{cases} 7 - 0.1 \times (25 - s_d), & \text{if } s_d < 25 \\ 7, & \text{if } 25 \leq s_d \leq 40 \\ 7 + 0.1 \times (s_d - 40), & \text{otherwise} \end{cases}$$

where s_d is the number of sentences in document d.

A *bracketed section* is defined as a group of terms where the leftmost and rightmost terms are significant terms, and no significant terms in the bracketed section are divided by more than four non-significant terms.

The *score for a bracketed section* is the square of the number of significant words falling in the section, divided by the total number of words in the entire sentence.

The *score for a sentence* is the maximum of all scores for the bracketed sections of the sentence.

Quoted from [4], technique based on [2], [3].