



Search Engine Overview

Ji-Rong Wen

WSM, MSRA
6/19/2006

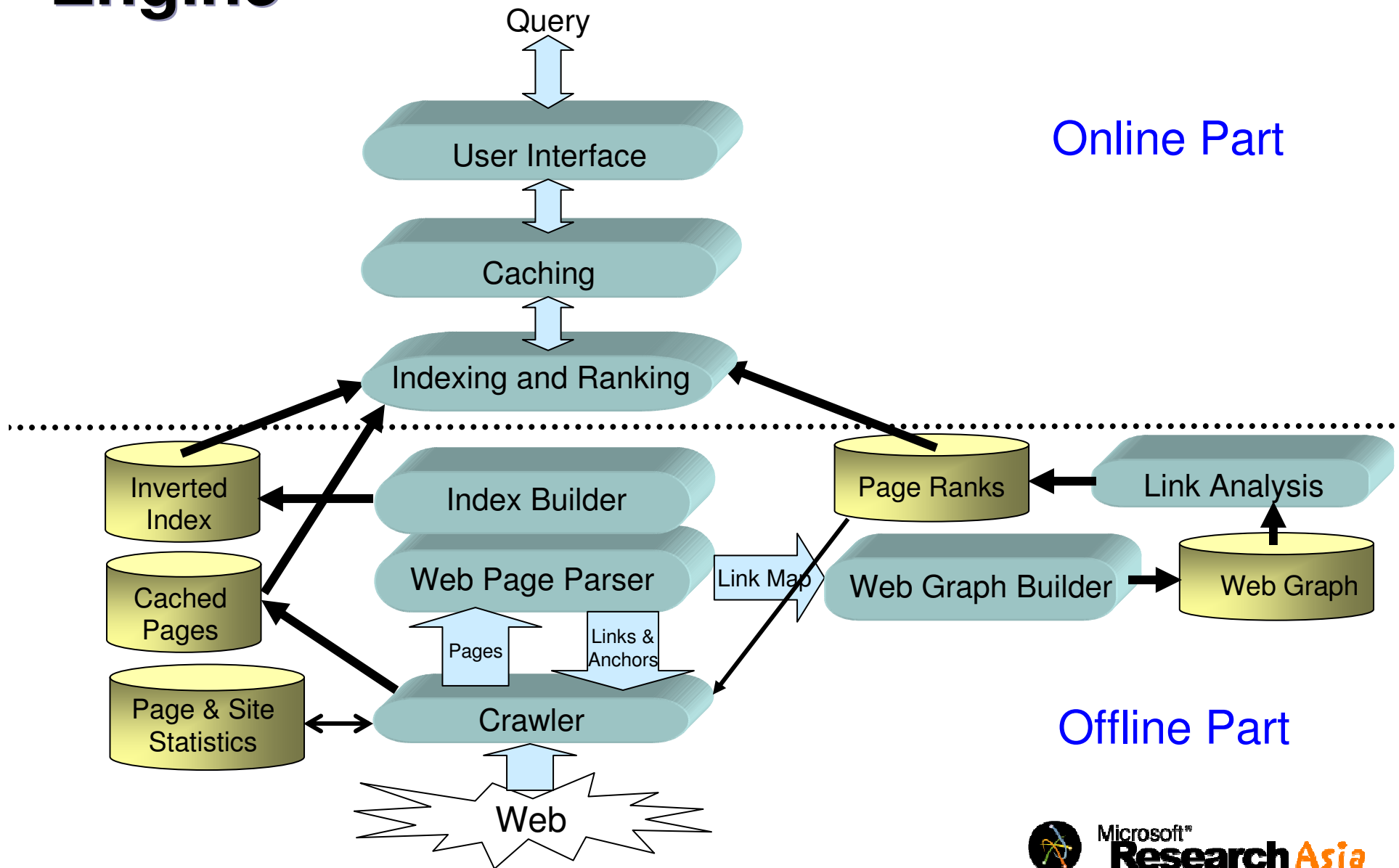
Outline

- A Simple Introduction to Search Engine Architecture
- Top 10 Challenges in Search Engine
- Top 10 Myths about Search Engine
- Computer Science in Search Engine

Outline

- **A Simple Introduction to Search Engine Architecture**
- Top 10 Challenges in Search Engine
- Top 10 Myths about Search Engine
- Computer Science in Search Engine

Architecture of a Typical Search Engine



Architecture – Crawler

➤ Functions

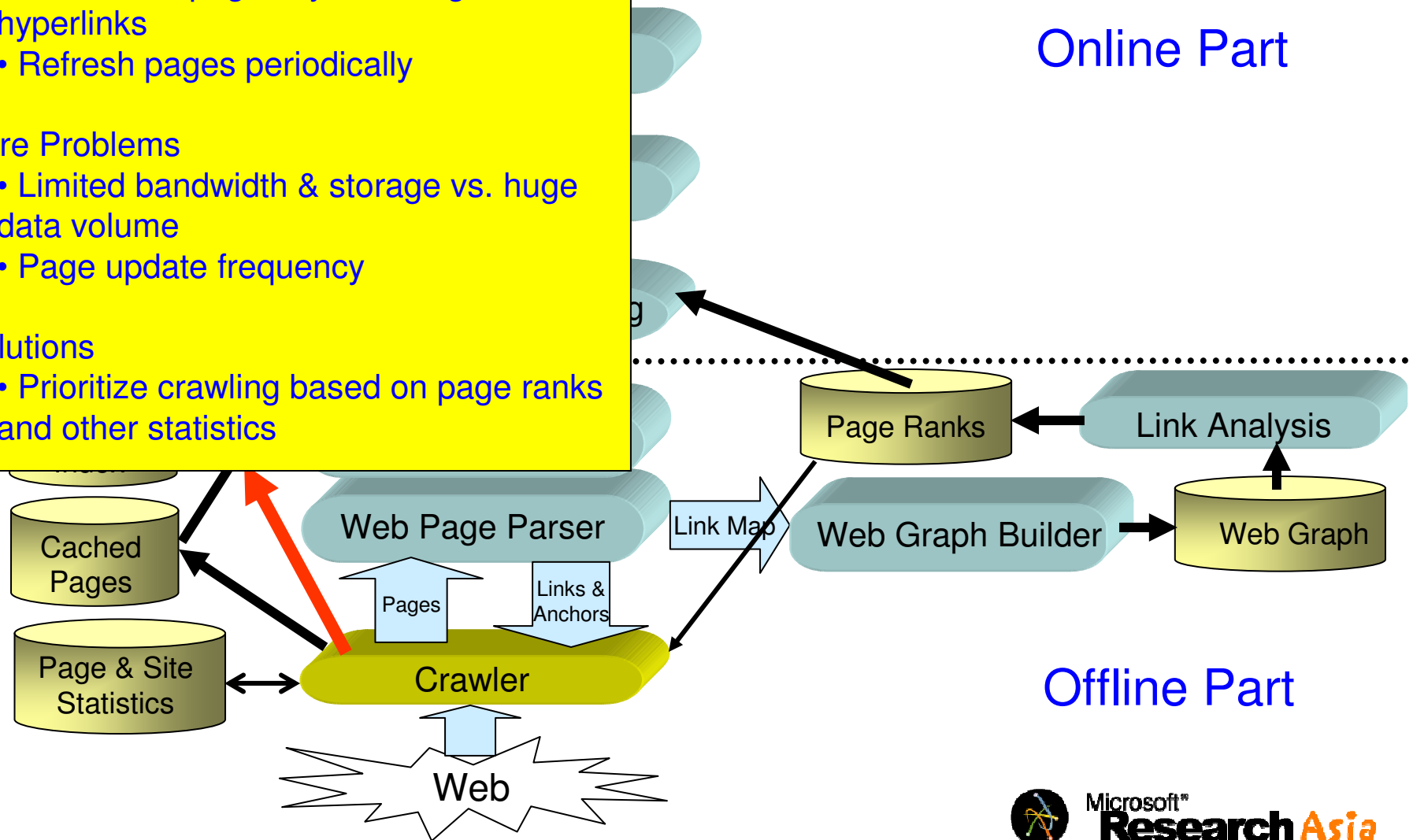
- Fetch Web pages by following hyperlinks
- Refresh pages periodically

➤ Core Problems

- Limited bandwidth & storage vs. huge data volume
- Page update frequency

➤ Solutions

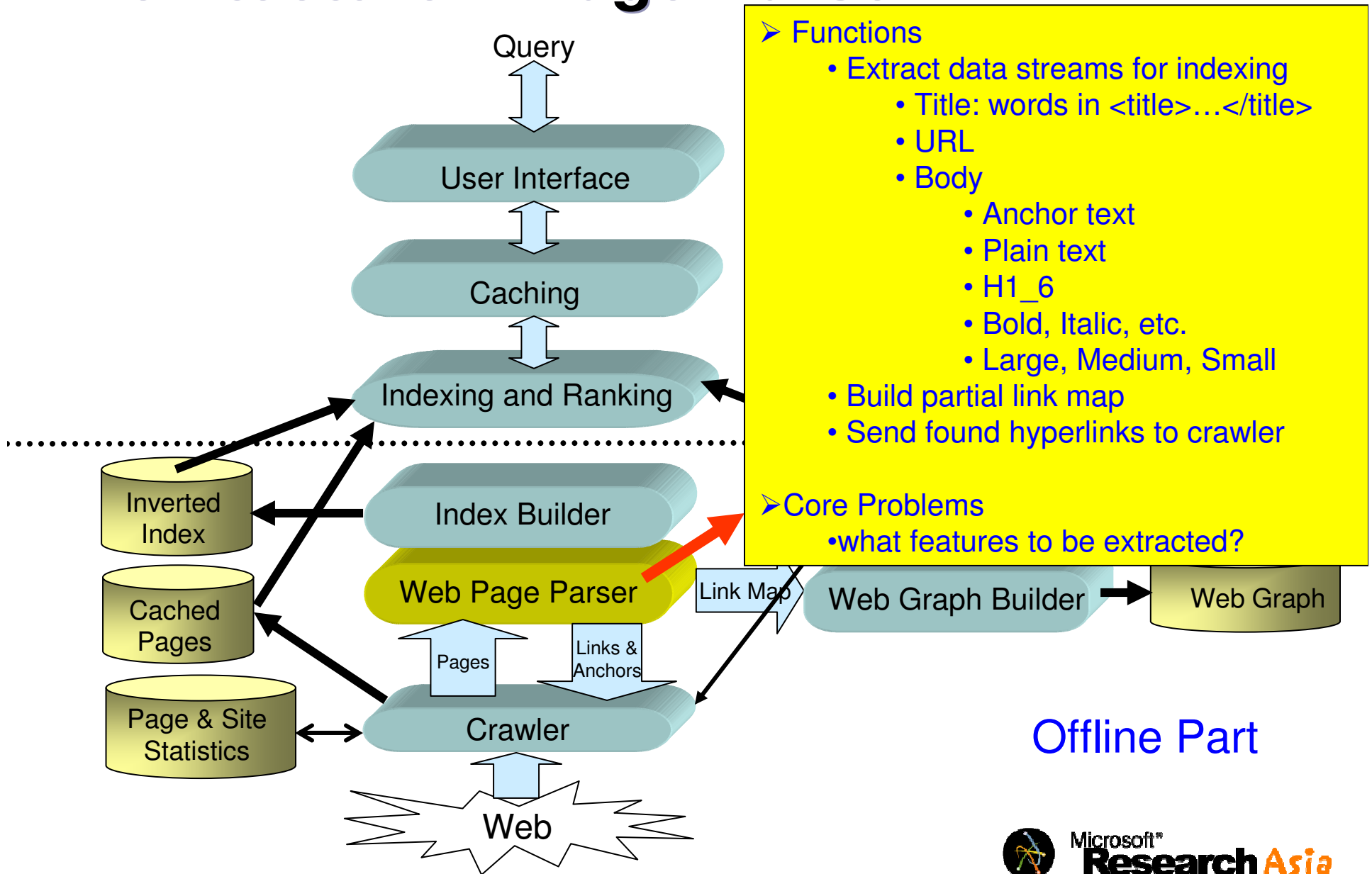
- Prioritize crawling based on page ranks and other statistics



Homework (1)

- How to estimate the refresh rate of a page?
- References
 - **Junghoo Cho, Hector Garcia-Molina. *Effective page refresh policies for Web crawlers.* ACM Transactions on Database Systems, 28(4): December 2003.**
 - **Junghoo Cho, Hector Garcia-Molina, Lawrence Page. *Efficient Crawling Through URL Ordering.* Computer Networks and ISDN Systems, 30(1-7):161-172, 1998**

Architecture – Page Parser

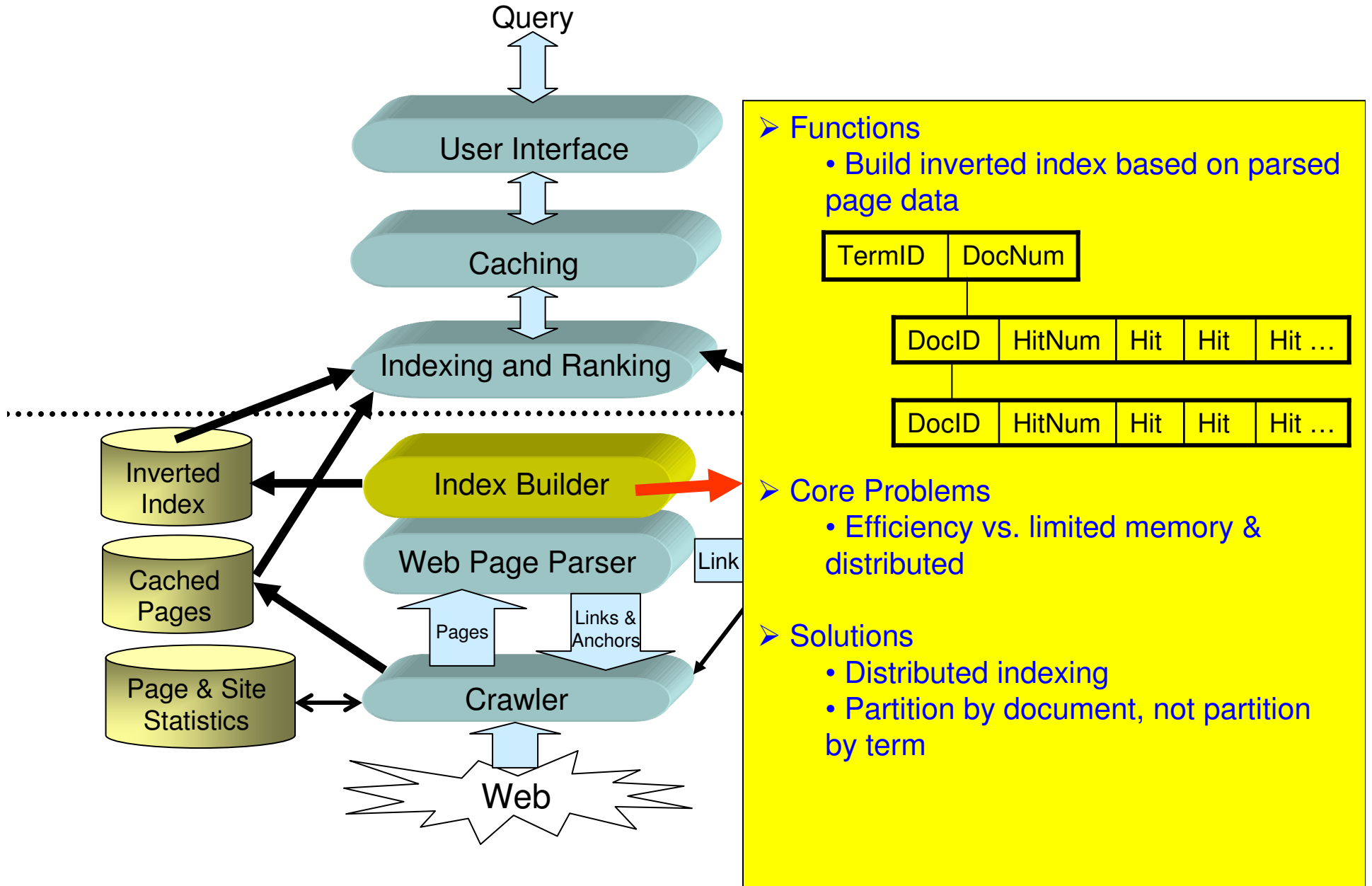


Offline Part

Homework (2)

- Write a Web page parser to get the terms in title, url, and body, with the position and font information for each term
- References
 - W3C HTML 4.01 Specification
<http://www.w3.org/TR/html4/>

Architecture – Index Builder



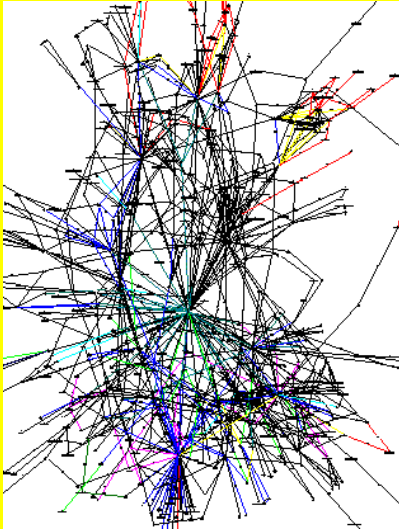
Homework (3)

- Write an inverted index building algorithm, with the following constraints:
 - a. memory is not sufficient to hold all documents
 - b. memory is not sufficient to hold the whole index
- References
 - Your “data structure” textbook

Architecture – Link Analysis *

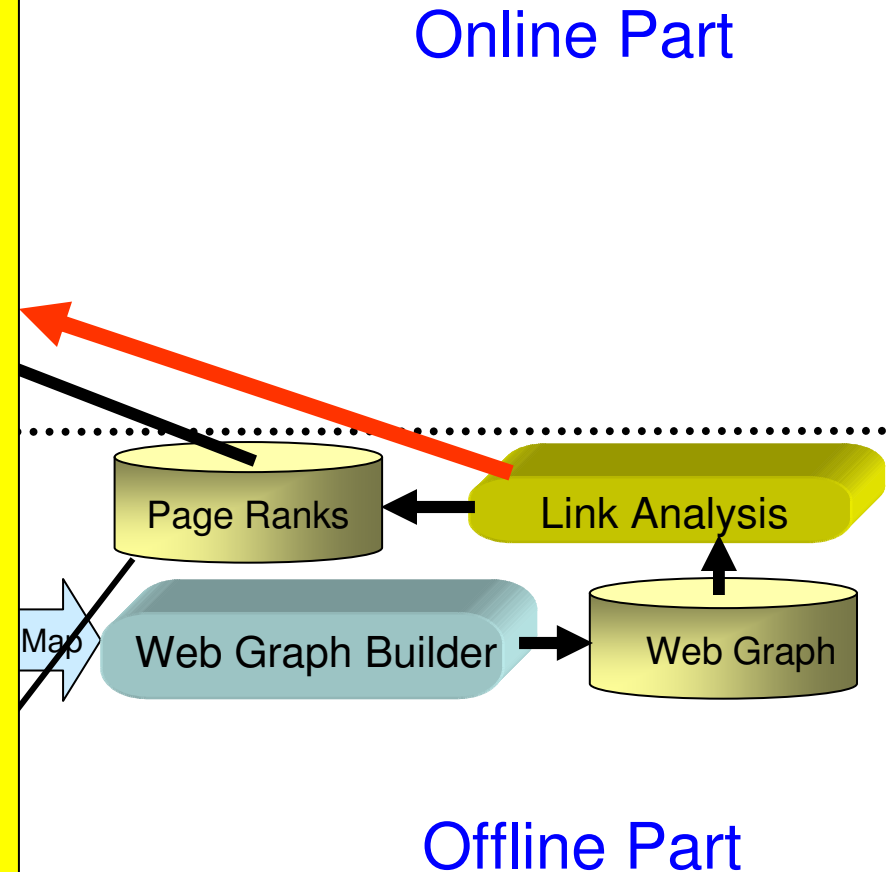
➤ Functions

- Measure the quality (or authority) of a page based on the link graph



➤ Core Problems

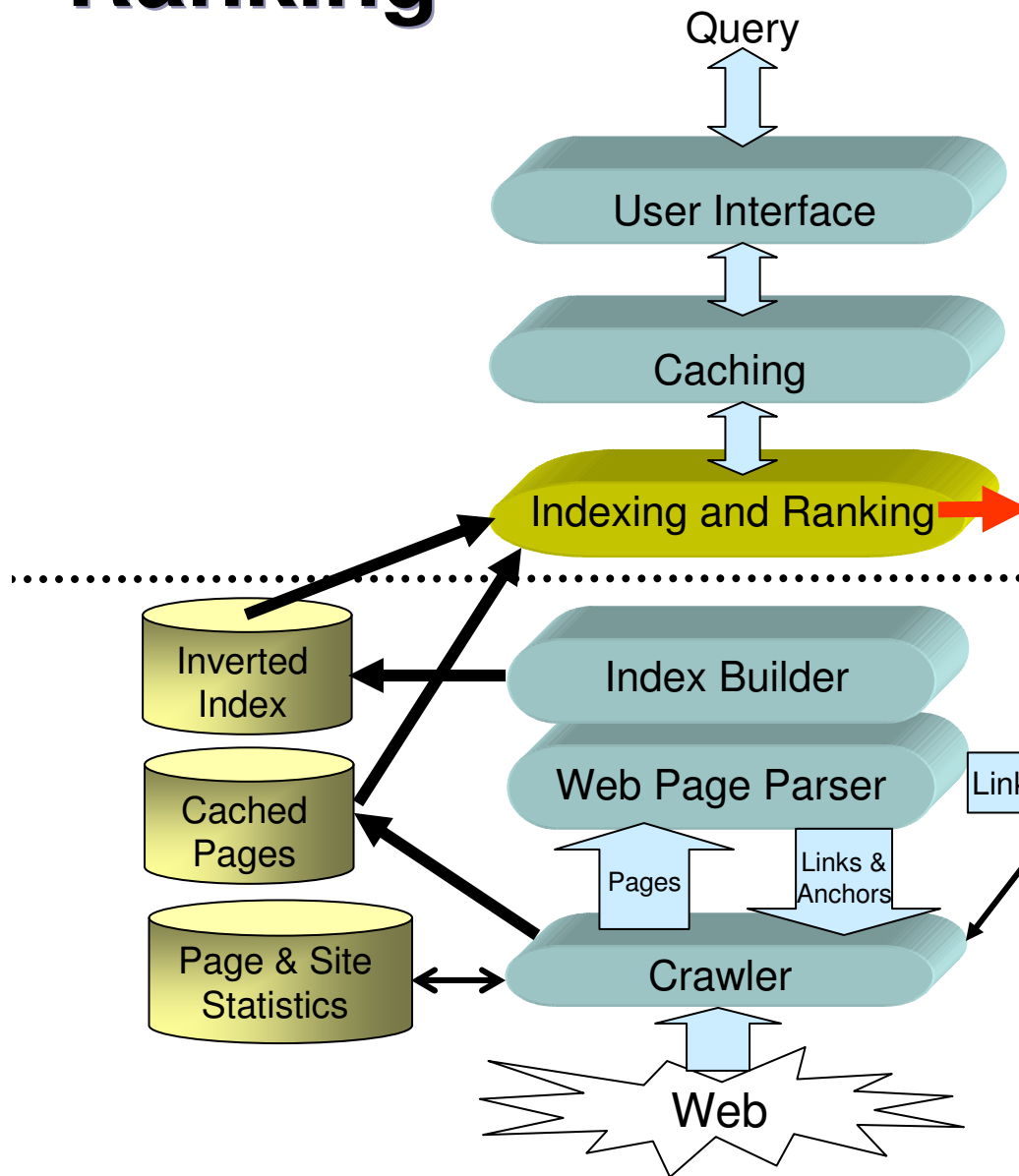
- Efficient algorithm on a huge graph
- Link-spam?
- Is link analysis the only way to determine the quality of pages?



Homework (4)

- Write a toy PageRank algorithm
- Why HITS algorithm is not a good choice for search engine?
- References
 - **Larry Page, Sergey Brin, R. Motwani, T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*** (1998), Stanford Digital Library Technologies Project
 - **Jon M. Kleinberg, *Authoritative Sources in a Hyperlinked Environment*** (1999), Journal of the ACM

Architecture – Indexing and Ranking *



➤ The core problems in the IR community, and has been studied for decades

➤ Functions

- Indexing: quickly locating pages containing query terms
- Ranking: sort pages according to relevance to the query

➤ Core Problems

- Performance: an inverted list for a hot term may be hundreds of megabytes.
- Accuracy: ranking functions with hundreds of parameters:
 - Anchor text
 - Page rank
 - Term proximity
 - $TF \cdot IDF$
 - ...

➤ Solutions

- Performance: Top-K query & index pruning
- Accuracy: tuning or learning?

Homework (5)

- Below is the slowest query I found on Google. Explain why. (*Hint: invalidating index pruning*)

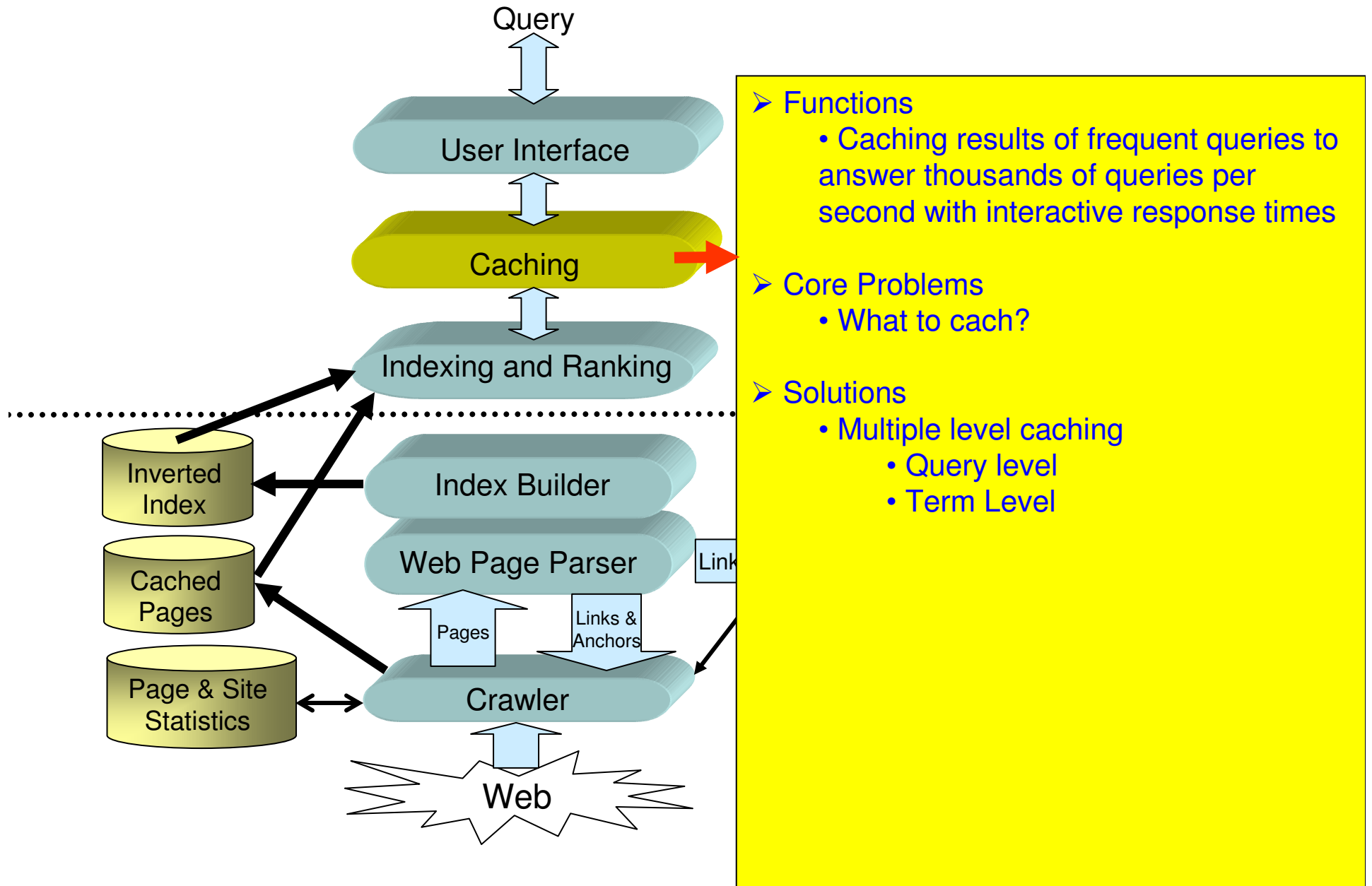
The screenshot shows a Google search interface. The search bar contains the query "a the the the is", which is circled in red. Below the search bar, there are radio buttons for "Search the Web" and "Search Chinese (Simplified) and Chinese (Traditional) pages". A message states: "There were no results in your selected language(s). Showing worldwide web results for 'a the the the is'." Below this, the search results are displayed under the heading "Web". The first result is "a the the the is" with a duration of "(14.52 seconds)", which is also circled in red. A red arrow points from the circled duration to the text "14.52 seconds" written in red below the screenshot. The search results include a tip about pinyin search, a software result for "EZ Address 1.4 .website content", and a result for "Thousands of free essays & papers on Your Bones in Space".

14.52 seconds

- References

- Xiaohui Long, Torsten Suel. *Optimized Query Execution in Large Search Engines with Global Page Ordering*. VLDB 2003

Architecture – Caching



Homework (6)

- Null
- References
 - Xiaohui Long, Torsten Suel. *Three-Level Caching for Efficient Query Processing in Large Web Search Engines*. 14th International World Wide Web Conference (WWW), 2005.



With the knowledge learned so far, you can build a decent single-machine search engine by yourself. Have a try if you want, it will only cost you several weeks of time.



But...

Some Facts of a Real Commercial Search Engine

- Huge data volume
 - 10B pages * 10K/per page = 100T
- Crawling bandwidth
 - $100T / (14 * 24 * 3600) = 82MB/second$
- Performance
 - 2000+ queries/second, response time < 1 second
- 10,000+ machines
 - System failure is normal: If one machine fails once in one year, $P(\text{at least one machine failed in each hour}) = 68\%$.
- High reliability: data are never allowed to corrupt
- High availability: 7*24 serving
- High scalability: machines are added or removed every day
- The electric power consumed by a large data center can supply a city with 50,000 people!

- The largest computer clusters in the world
 - A lot of tough things to be solved



Outline

- A Simple Introduction to Search Engine Architecture
- **Top 10 Challenges in Search Engine (排名不分先后)**
- Top 10 Myths about Search Engine
- Computer Science in Search Engine

#1: Spamming *

- Click → Money, Spam → Click ==> Spam → Money
- An endless game between spammers and search engines



Web

News results for [miserable failure](#) - [View today's top stories](#)



[Failure's not an option](#) - Indianapolis Star - 14 Jun 2006

[Biography of President George W. Bush](#)

Biography of the 43rd President of the United States.

www.whitehouse.gov/president/gwbbio.html - 26k - [Cached](#) - [Similar pages](#)

[Biography of Jimmy Carter](#)

Short biography from the official White House site.

www.whitehouse.gov/history/presidents/jc39.html - 36k - [Cached](#) - [Similar pages](#)

Homework (7)

- Prove either of the following propositions:
 - There are spam-immune ranking algorithms
 - There is NO a spam-immune ranking algorithm

#2: Data Acquisition

- Growing speed of the Web >> Growth of indexing capability of search engines.
- Re-crawl frequently updated pages: news, blog, bbs
- Dynamic contents: deep Web, Web 2.0
- Crawling is the first step of search, but its importance is largely ignored by academia.

Homework (8)

- How to crawl blogs?

#3: Content Quality

- Traditional IR: an assumption is that every document in a collection is authoritative and accurate.
- Link analysis is one way, anything else? How about cases lacking of links?
 - User clicks
 - Absolute link number vs. relative link growth
 - Page organization
 -

#4: Ranking *

- No More Things to Say 😊

#5: Evaluation *

- Traditional IR evaluation
 - Limited binary judgment
 - Small query set
 - Small and static document collection
- Web Evaluation
 - Result quality is important
 - Multiple level judgment
 - Query distribution is changing
 - Pages are changing consistently
- Clicks as implicit judgment?

#6: Query Formulation

- How do you compose your queries?
 - Guess if the terms occur in the wanted pages
 - Relevant to terms, instead of relevant to query
 - What to do if the guess fails?

The image shows two side-by-side screenshots of a Google search interface. The left screenshot shows the search results for the query 'lei zhang'. The right screenshot shows the search results for the query '(86-10) 62617711 ext. 3197'. Both screenshots have red circles highlighting specific search results.

Left Screenshot (Query: lei zhang):

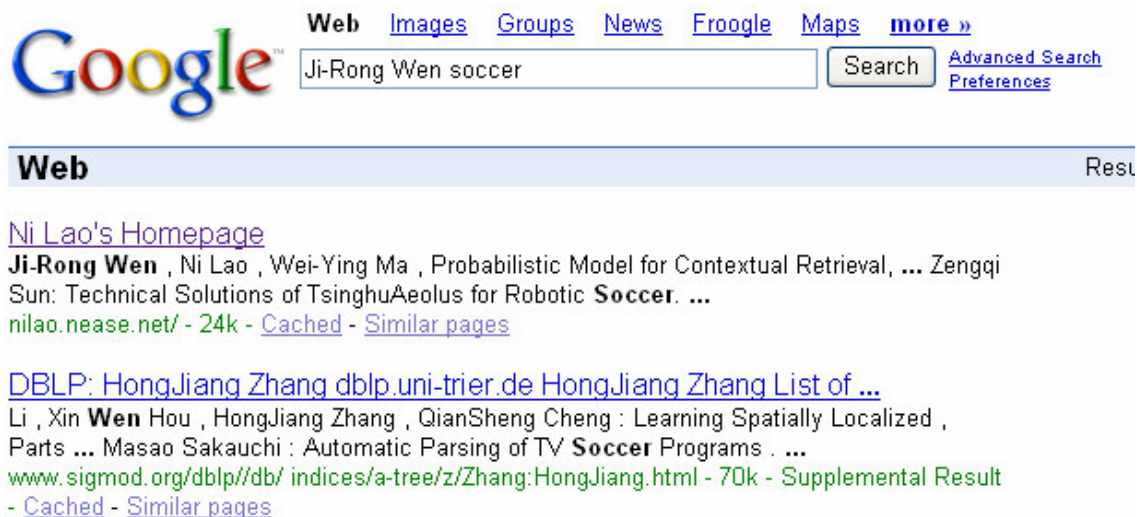
- Search bar: lei zhang
- Results:
 - [Welcome to Zhang Lei's Homepage](#)
www.acm.caltech.edu/~zhanglei/ - 1k - 17 Jun 2006 - [Cached](#) - [Similar pages](#)
 - [Lei Zhang's Homepage](#)
Lei Zhang. Voice: 631-632-8427. Department of Computer Science. URL: http://www.cs.sunysb.edu/~lzhang. Stony Brook, NY, 11794, USA ...
www.cs.sunysb.edu/~lzhang/ - 45k - [Cached](#) - [Similar pages](#)
 - [Lei Zhang's Home Page](#)
Lei Zhang is an associate researcher in Microsoft Research Asia.
research.microsoft.com/users/leizhang/ - 21k - [Cached](#) - [Similar pages](#)

Right Screenshot (Query: (86-10) 62617711 ext. 3197):

- Search bar: (86-10) 62617711 ext. 3197
- Results:
 - [Lei Zhang's Home Page](#)
No. 49, Zhichun Road, Haidian District, Beijing 100080, PRChina. Email: leizhang AT microsoft.com. Tel: (86-10) 62617711 ext. 3197. Fax: (86-10) 62555337 ...
research.microsoft.com/users/leizhang/ - 21k - [Cached](#) - [Similar pages](#)
 - [Michelle Aalbers Laura Abba Phone: +39 050 3152633 laura.abba ...](#)
... 972 916 4292 steven.dekany @ marconi.com Alan DeKok Phone: 613 724 6004 ext. ...
cs.ucla.edu Qian Zhang Phone: 86-10-62617711-3135 qianz @ microsoft.com ...
www3.ietf.org/proceedings/01dec/atts.txt - 114k - [Cached](#) - [Similar pages](#)

#7: Personalization

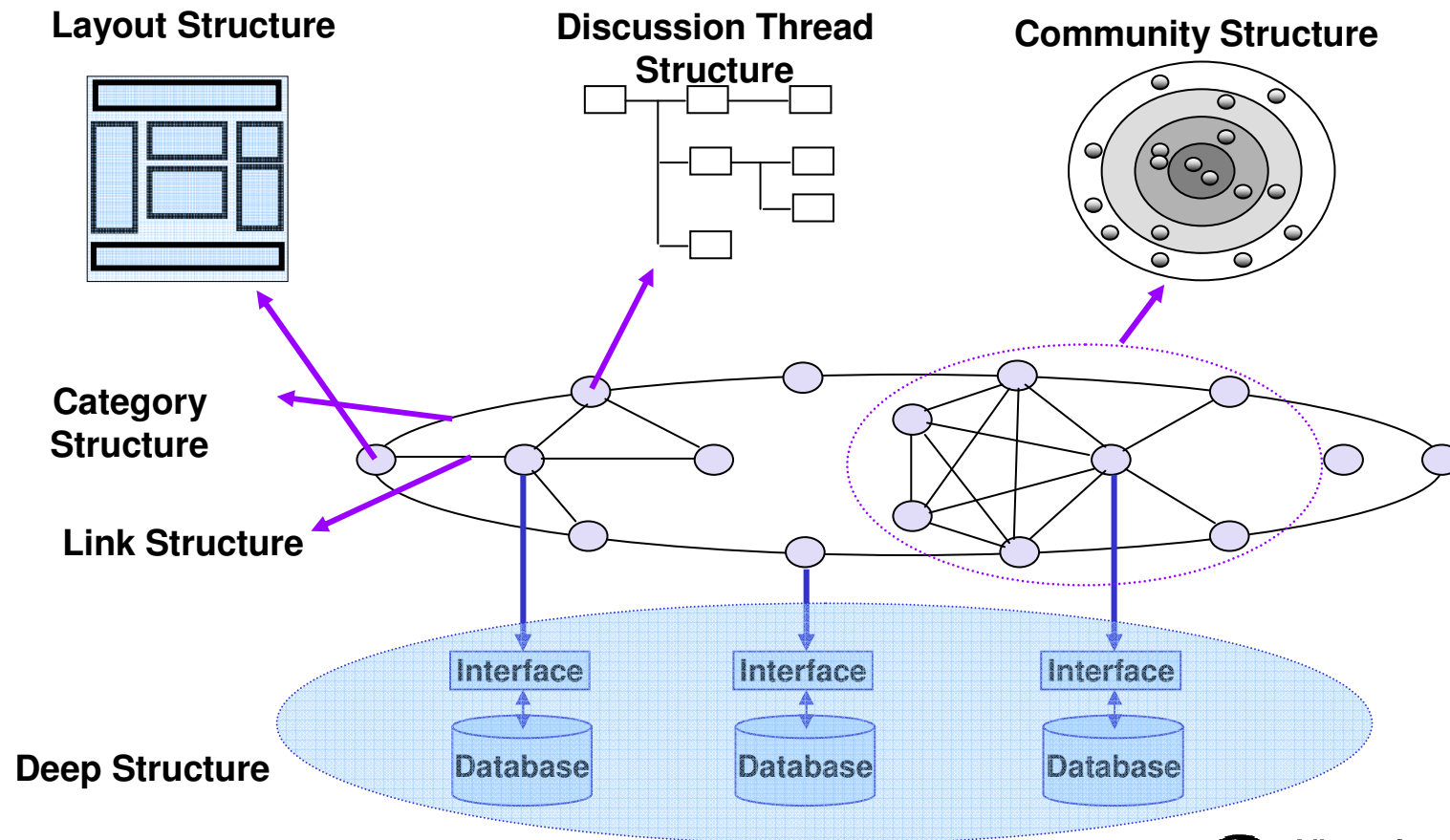
- Personalized search, a long history, but never a success story
- Do you really need personalization?
- When do you need personalization?
- When needed, can you easily find an alternative way?
- When to apply personalization?



The screenshot shows a Google search interface. At the top left is the Google logo. To its right are navigation links: Web, Images, Groups, News, Froogle, Maps, and more ». Below these is a search input field containing the text 'Ji-Rong Wen soccer' and a 'Search' button. To the right of the search button are links for 'Advanced Search' and 'Preferences'. Below the search bar is a horizontal bar with the word 'Web' on the left and 'Rest' on the right. The search results are listed below this bar. The first result is 'Ni Lao's Homepage' with a sub-line: 'Ji-Rong Wen , Ni Lao , Wei-Ying Ma , Probabilistic Model for Contextual Retrieval, ... Zengqi Sun: Technical Solutions of TsinghuaAeolus for Robotic Soccer. ... nilao.nease.net/ - 24k - Cached - Similar pages'. The second result is 'DBLP: HongJiang Zhang dblp.uni-trier.de HongJiang Zhang List of ...' with a sub-line: 'Li , Xin Wen Hou , HongJiang Zhang , QianSheng Cheng : Learning Spatially Localized , Parts ... Masao Sakauchi : Automatic Parsing of TV Soccer Programs www.sigmod.org/dblp/db/ indices/a-tree/z/Zhang:HongJiang.html - 70k - Supplemental Result - Cached - Similar pages'.

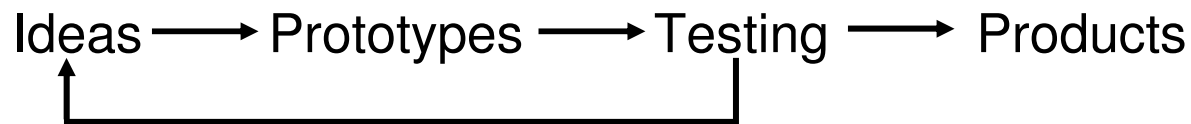
#8: Structure in the Web *

- Are Web data really unstructured?
- More structure = better search



#9: Infrastructure *

- The cycle of Web innovation



- A platform for creating high-quality products
- A platform for quick innovations
 - How difficult to test a new algorithm in 5B pages?
 - How difficult to calculate the query frequencies in 100T search logs?

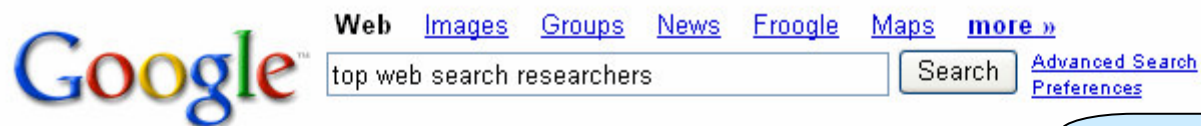
#10: The Next Big Thing?

Outline

- A Simple Introduction to Search Engine Architecture
- Top 10 Challenges in Search Engine
- **Top 10 Myths about Search Engine (排名不分先后)**
- Computer Science in Search Engine

#1

- **Myth:** Some search engines are close to “perfect”.
- **Fact:** They are perfect because you have no choice
 - Search engines lower our expectations
 - We are getting used to their poor performance



Web

[Choose the Best Search for Your Information Need](#)

Library and Archival Exhibitions on the **Web**, **Search** for online exhibitions selected by Smithsonian ... I need to do **research** in a specific discipline... ...

www.noodletools.com/debbie/literacies/information/5locate/adviceengine.html - 83k - [Cached](#) - [Similar pages](#)

[Library of Congress WWW/Z39.50 Gateway](#)

Contents: **Search** Library of Congress Catalog | **Search** Other Catalogs | About the ... Commonwealth Scientific and Industrial **Research** Organization (CSIRO) ...

www.loc.gov/z3950/ - 94k - [Cached](#) - [Similar pages](#)

Is this perfect?

#2

- **Myth:** There are magic algorithms in search engines
- **Fact:** There is no a single magic algorithm can make you win the search battle
 - PageRank is not that important as you think. It is only one small factor among many many others that search engines use to determine the ranking
 - Search algorithms are keeping improving

#3

- **Myth:** Most of the information on the Web has been indexed by search engines.
- **Fact:** Only a very tiny fraction of Web information is being indexed.
 - Seen URLs >> crawled URLs
 - Dynamic contents, deep Web, Web 2.0 contents

#4

- **Myth:** It is easy to switch to another search engine.
- **Fact:** Users only switch to a search engine significant better than the current one.

#5

- **Myth:** Ranking is the most important thing
- **Fact:** An infrastructure enabling quick innovations is most important
 - No good infrastructure, no good ranking
 - Good ranking is the result of many hard efforts behind

#6

- **Myth:** Search engine is equivalent to Web information retrieval
- **Fact:** Search engine is equivalent to Web-scale information *management*
 - Information acquisition, processing, storage, access, indexing, querying, mining

#7

- **Myth:** Cool feature is the king
- **Fact:** Do “simple” thing and do it best is the king
 - In terms of features, the current search engines are in fact the same as those ten years ago
 - Ideas vs. ideas do work!
 - Of course, only if you have a really cool idea that can change the game

#8

- **Myth:** Ideas in top conference papers are excellent
- **Fact:** Many of them DO NOT work at all!
 - Toy system
 - Small dataset
 - Scholastic evaluation
 - We are changing this sad truth!

#9

- **Myth:** Most of Web search researchers are from the IR community
- **Fact:** They come from diverse fields
 - Researcher in WSM group are from multimedia, database, machine learning, system, IR, etc.

#10

Outline

- A Simple Introduction to Search Engine Architecture
- Top 10 Challenges in Search Engine
- Top 10 Myths about Search Engine
- **Computer Science in Search Engine**

Information Retrieval in SE

- Information Retrieval \neq Text Retrieval
- Information Retrieval = Information Retrieval
 - Web is the largest information source
- Go to check the percentage of Web search related papers in SIGIR'98 – SIGIR'06

Systems in SE

- Search engine data centers: the largest distributed computing platforms in the world
 - When the scale is large enough, it becomes a system problem 😊
- Infrastructure for Web-scale data processing
 - is it Web OS?

Database in SE

- Is Web a Huge Database?
 - Most data on the Web are in fact (semi-)structured
 - Database people want to manage more data 😊
- Online Database everywhere
- “DB+IR” workshops in SIGMOD, VLDB, SIGIR, and WWW
- “WebDB” workshop
- WebDB? a long way to go...

	<i>Database (DR)</i>	<i>Information Retrieval (IR)</i>
<i>Data</i>	Structured	Unstructured
<i>Model</i>	Deterministic	Probabilistic
<i>Inference</i>	Deduction	Induction
<i>Query language</i>	Artificial	Natural
<i>Query specification</i>	Complete	Incomplete
<i>Matching</i>	Exact match	Partial match, best match
<i>Items wanted</i>	Matching	Relevant
<i>Error response</i>	Sensitive	Insensitive
<i>Data update</i>	Full-support	Not support
<i>Transaction</i>	Support	Not support
<i>Usage</i>	Application-oriented	Human-oriented

Machine Learning & Data Mining in SE

- Data! A huge amount of data!! Various kinds of data!!!
 - Data mining and machine learning people are exciting...
 - *"If you have a lot of data, then you don't need a lot of methodology."*
- All Web-scale data processing tasks needs to be automatic
- Learning to
 - Learning to ranking *
 - Learning to crawling
 - Learning to extracting *
 -

Others

- Multimedia
 - Social Science
 - User Interface
 - Network
 - Hardware
 -
-
- You can get a PhD degree by working on Web search problems☺



Thanks!