Mining di Dati Web

Lesson no. 1

Course Description

- Teachers: Dino Pedreschi[*] e Fabrizio Silvestri[+]

- Dipartimento di Informatica,
  Università di Pisa
  Via Buonarroti, 2
  I-56125 Pisa
  Italy
  email: pedre@di.unipi.it
  web: http://www.di.unipi.it/~pedre/

- [+]ISTI - CNR
  Area della Ricerca di Pisa
  Via G. Moruzzi, 1
  Room C-34
  Phone +39 050 315 3011
  email: fabrizio.silvestri@isti.cnr.it
  web: http://hpc.isti.cnr.it/~silvestr

Introduction

- Outline of the course
- What’s the Web
- Modern Web Search Engines
- Mathematical Web Models

HOMER'S WORLD
SKY CINE NEWS

What's the Web?

- Hypertext documents
  - Text
  - Links
- Web
  - Billions of documents
  - Authored by millions of diverse people
  - Edited by no one in particular
  - Distributed over millions of computers, connected by variety of media

Some Historical Notes

- Memex, 1945 [Vannevar Bush, US President Roosevelt's science advisor]
  - Stands for "MEMory EXtension"
  - Aim: to create and help follow hyperlinks across documents
  - Photoelectrical-mechanical storage and computing device that could store vast amounts of information, in which a user had the ability to create links of related text and illustrations. This trail could then be stored and used for future reference. Bush believed that using this associative method of information gathering was not only practical in its own right, but was closer to the way the mind ordered information."

The Term Hypertext

- Hypertext, term coined by Ted Nelson in a 1965 paper to the ACM 20th national conference:
  - [...] By 'hypertext' mean nonsequential writing - text that branches and allows choice to the reader, best read at an interactive screen

Hypertext Systems

- The first hypertext-based system was developed in 1967 by a team of researchers led by Dr. Andries van Dam at Brown University
- The research was funded by IBM and the first hypertext implementation, Hypertext Editing System, ran on an IBM/360 mainframe
- IBM later sold the system to the Houston Manned Spacecraft Center which reportedly used it for the Apollo space program documentation

Hypertext Systems

- Xanadu hypertext, by Ted Nelson, 1981:
- In the Xanadu scheme, a universal document database (docuverse), would allow addressing of any substring of any document from any other document. "This requires an even stronger addressing scheme than the Universal Resource Locators used in the World-Wide Web." [De Bra]
- Additionally, Xanadu would permanently keep every version of every document, thereby eliminating the possibility of a broken link. Xanadu would only maintain the current version of the document in its entirety

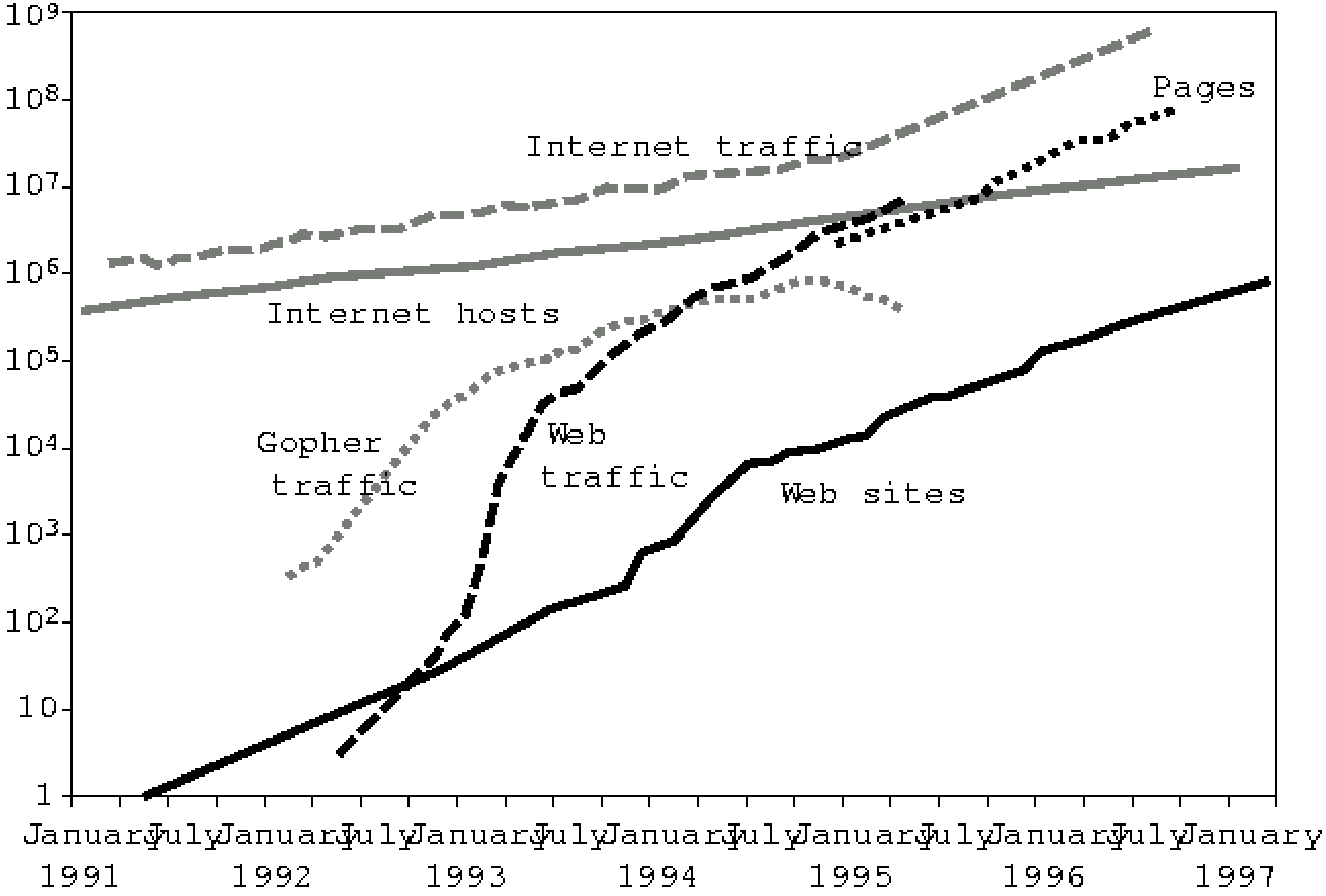
Hypertext Systems

- World Wide Web
- Initiated at CERN in 1989 by Tim Berners-Lee, now w3c director:
- “W3 was originally developed to allow information sharing within internationally dispersed teams, and the dissemination of information by support groups. Originally aimed at the High Energy Physics community, it has spread to other areas and attracted much interest in user support, resource discovery and collaborative work areas. It is currently the most advanced information system deployed on the Internet, and embraces within its data model most information in previous networked information systems.”

WWW User Interfaces

- GUIs
  - Berners-Lee (WorldWideWeb - 1990)
  - Erwise and Viola(1992), Midas (1993)
- Mosaic (1993)
  - A hypertext GUI for the X-window system
  - HTML: markup language for rendering hypertext
  - HTTP: hypertext transport protocol for sending HTML and other data over the Internet
  - CERN HTTPD: server of hypertext documents

10 years ago: Some Numbers



1994: The Landmark Year

- Foundation of the “Mosaic Communications Corporation” (later Nestcape)
- First World-Wide Web conference
- MIT and CERN agreed to set up the World-wide Web Consortium (W3C)

Web: A populist, participatory medium

- Number of writers =(approx) number of readers
- The evolution of MEMES (Richard Dawkins):
    - Memes=genes - ideas, theories etc that spread from person to person by imitation.
    - Now they have constructed the Internet!!
    - E.g.: "Free speech online", chain letters, and email viruses

Abundance and Authority Crisis

- Liberal and informal culture of content generation and dissemination
- Very little uniform civil code
- Redundancy and non-standard form and content
- Millions of qualifying pages for most broad queries
  - Example: java or kayaking
- No per se authoritative information about the reliability of a site

Problems with Uniform Accessibility

- Little support for adapting to the background of specific users.
- Commercial interests routinely influence the operation of Web search
- Users pay for connection costs, not for contents
- Profit depends from ads, sales, etc
- SEO: “Search Engine Optimization“!!

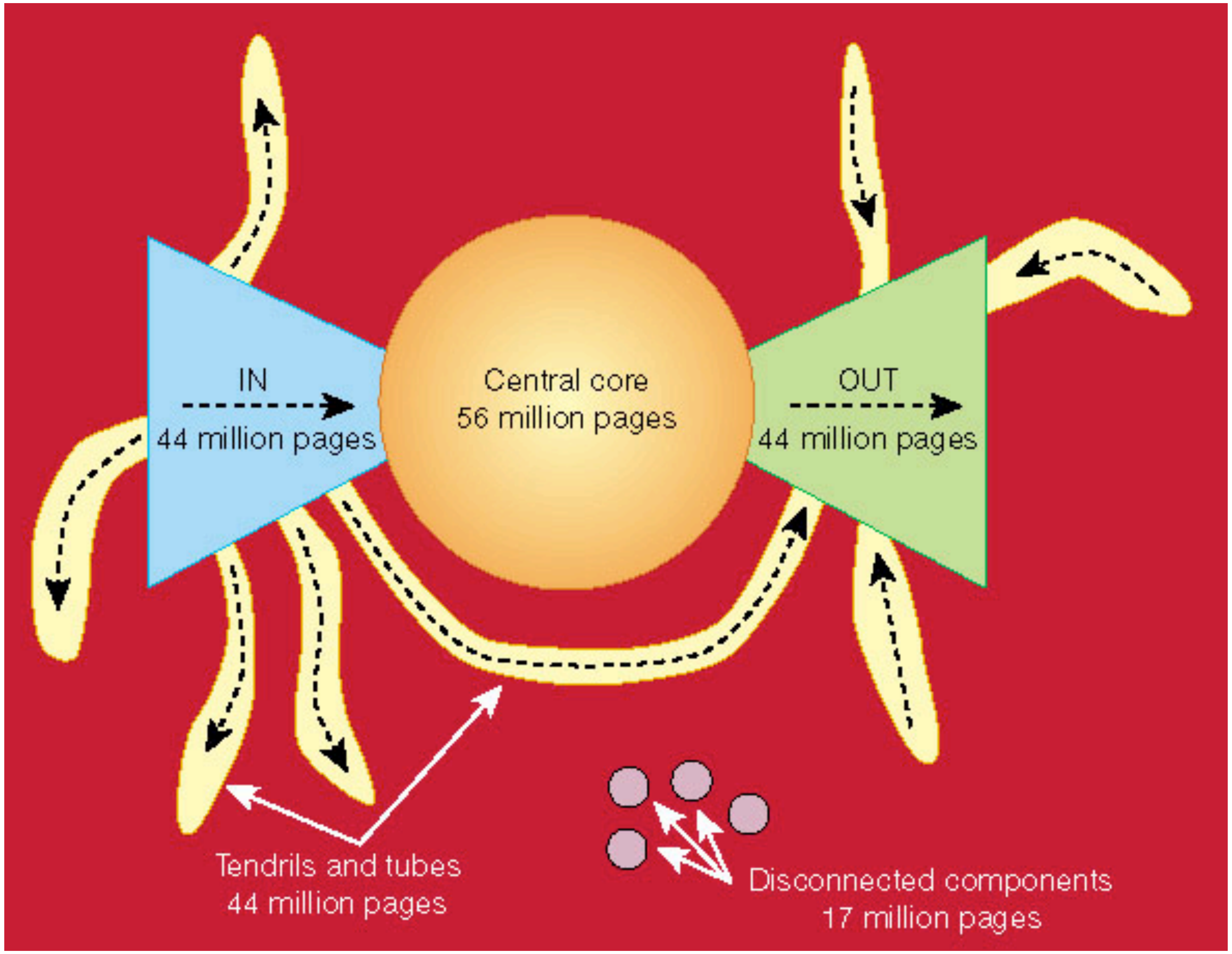
WWW Data

- Semi-structured or unstructured
  - No DB schema!
- Large number of attributes
  - Each word can be considered an attribute

Data vs. Web Mining

- Traditional data mining
  - data is structured and relational
  - well-defined tables, columns, rows, keys, and constraints
- Web data
  - readily available data rich in features and patterns
  - spontaneous formation and evolution of
  - topic-induced graph clusters
  - hyperlink-induced communities
  - Web Mining: discovering patterns which are spontaneously driven by semantics

A Picture is Worth Thousand Words



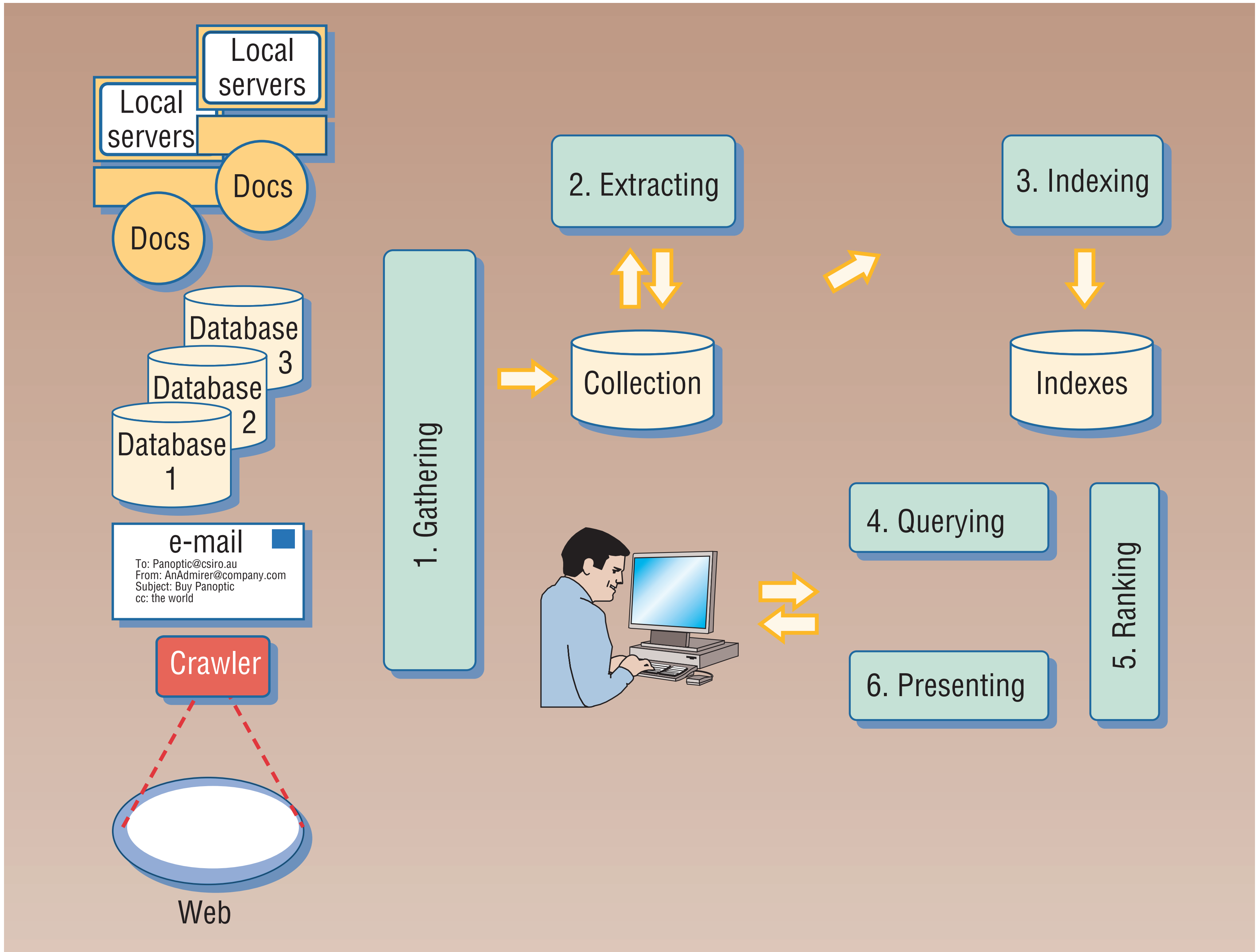
Source: The web is a bow tie. Nature 405, 113(11 May 2000)

Web Search

Based on: R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, F. Silvestri: Challenges on Distributed Web Retrieval. ICDE 2007: 6-20

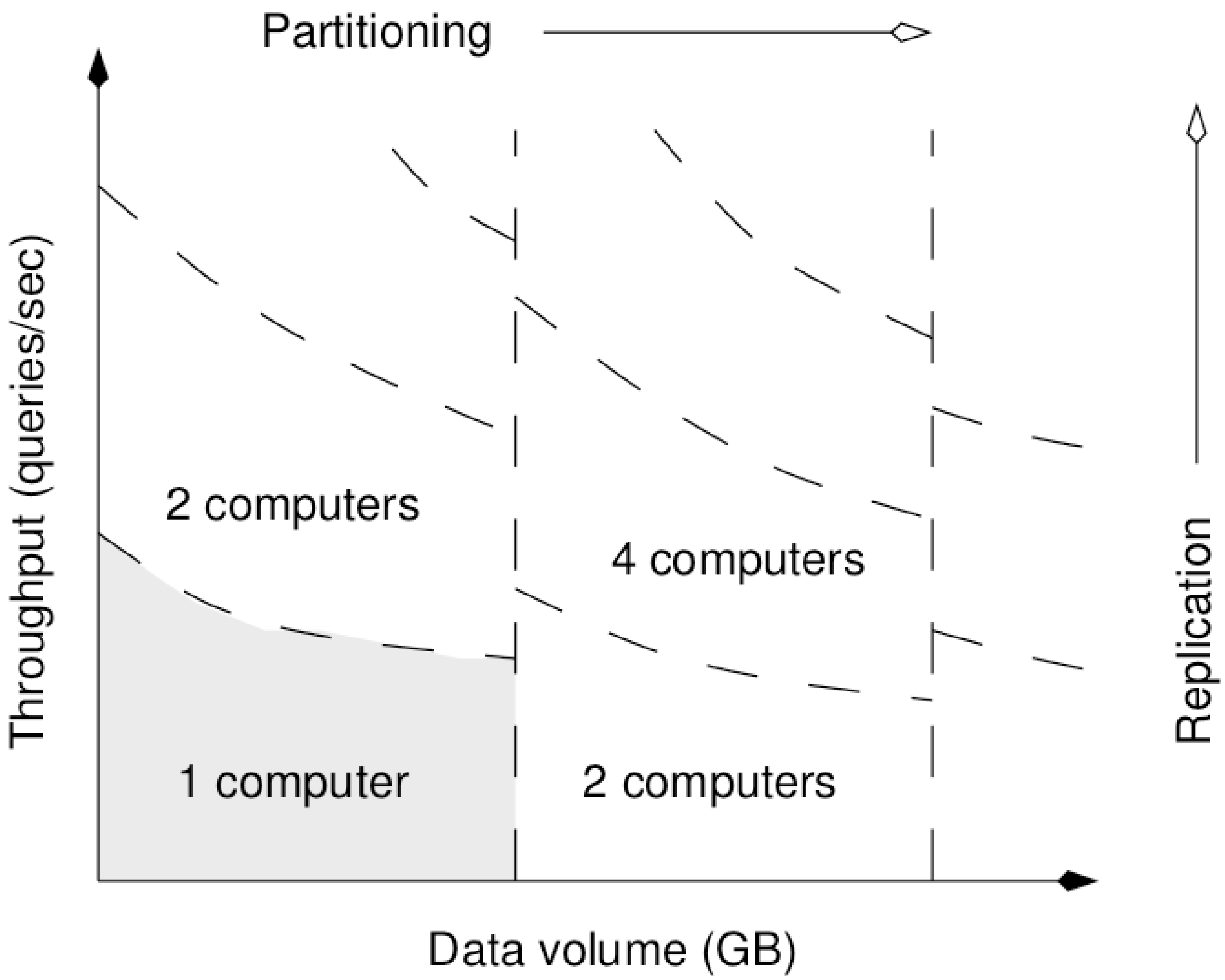
- It is, without any doubts, the most complex data engineering task today:
  - Distributed in nature
  - Large volume of data
  - Highly concurrent service
- Today Web Search Engines are organized as Large Replicated Centralized Architecture.

What Centralized Architecture?



David Hawking: Web Search Engines: Part 1. IEEE Computer 39(6): 86-88 (2006)

Large Replicated Centralized Architecture



A. Moffat, J. Zobel: What Does It Mean to "Measure Performance"? WISE 2004: 1-12

Out-of-the-Envelope Computation

- 20 billion web page implies, at least, 100Tb of text
- The index in RAM implies at least a cluster of 3,000 PCs
- Assume we can answer 1,000 queries per second
- 173 million queries a day imply 2,000 queries per second
- Decide that the peak load plus a fault tolerant margin is 5 times the average
- This load implies a replication factor of 10 giving a total of 30,000 PCs
- Total deployment cost of over 100 million US$ plus maintenance cost

Being conservative, in 2010, we will need over 1 million computers!

Challenges

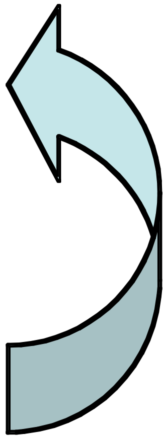
- A Search Engine:
  - Must return high quality results (e.g. handle quality diversity, and fight spam)
  - Must be fast (fraction of a second)
  - Must have high capacity
  - Must be dependable (reliability, availability, safety, and security)
  - Must be scalable

Main Modules & Issues

	Partition	**Dependability**	**Communication (sync.)**	**External factors**
Crawling	URL assignment	Re-crawl	URL exchanges	Web growth, Content change, Network topology, Bandwidth, DNS, QoS of servers
Indexing	Doc. partition, Term partition	Re-index	Partial indexing, updating, merging	Web growth, Content change, Global statistics
Querying	Query routing, Collection selection, Load balancing	Replication, caching	Rank aggregation, Personalization	Changing user needs, User base growth, DNS

Crawler

- It "simply" does the *crawling*
- In theory it is simple:
  - fetch a Web page
  - parse it
  - extract links



- Actually, it implies using other people's resources: Web Servers' CPU cycles, and bandwidth

Issues

- How to partition the crawling task?
- What to do when one agent fails?
- How to communicate among agents?
- How to deal with external factors?

Partition

- Host-based partitioning to exploit locality of links
- Balance improves if large/small hosts are treated differently
- Performance increases if geographic location is considered:
    - no research results, only industry seems to follow this path

Communication

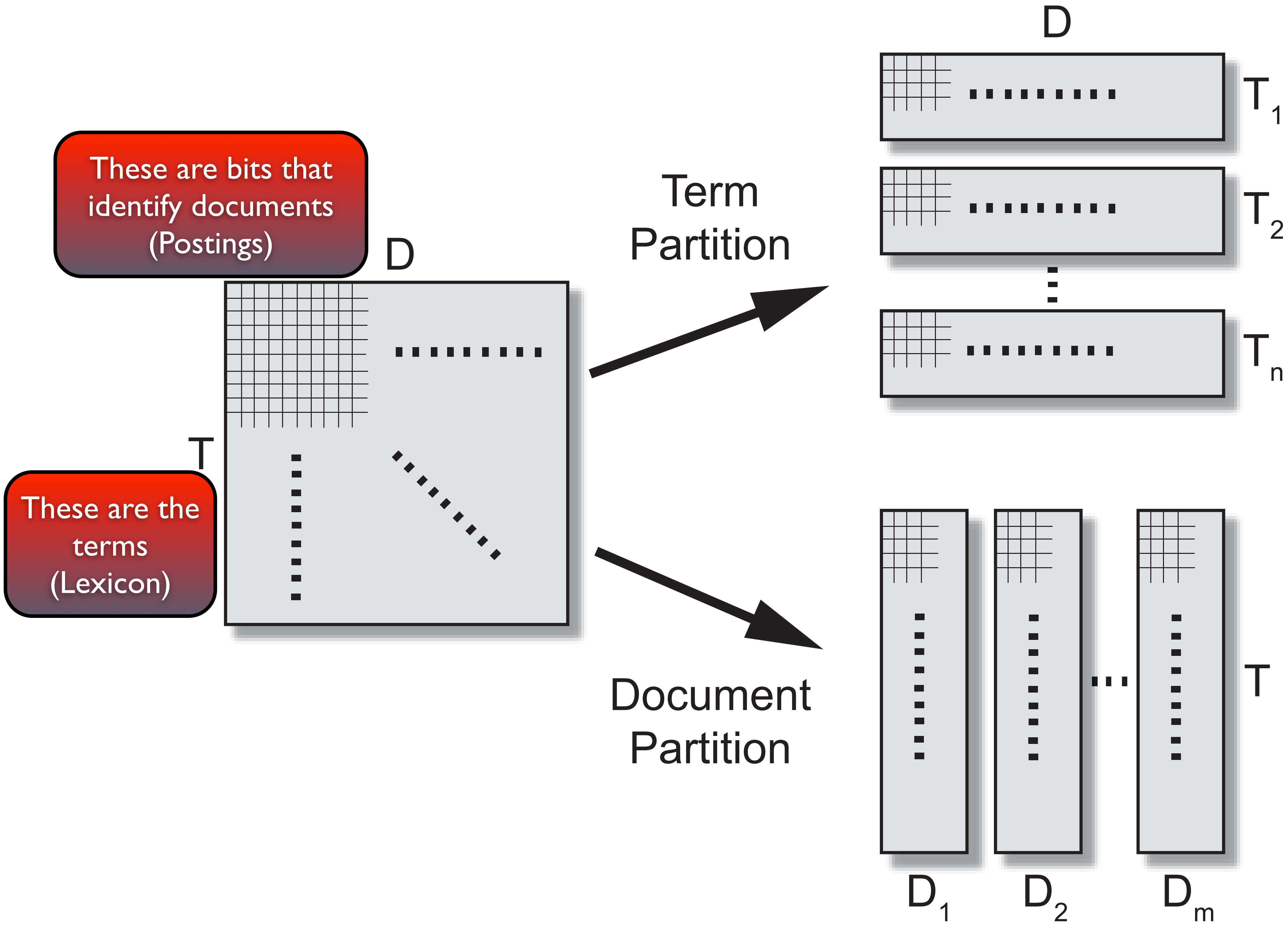
- Host-based partitioning reduces communications because of the locality of the links
- Highly linked URLs should be not crawled repeatedly
- Communication with the server can be improved if server cooperates:
  - O. Brandman, J. Cho, H. Garcia-Molina, N. Shivakumar: Crawler-Friendly Web Servers. SIGMETRICS Performance Evaluation Review 28(2): 9-14 (2000)

Indexing

- Indexing in Database and IR is the process of building an *index* over a stored collection of data
- In IR a structure called *Inverted Index* is usually employed
  - *Lexicon*: contains distinct terms appearing in the collection's documents
  - *Posting Lists*: contains descriptions of occurrences of relative terms within the corresponding documents

Web Search Engines are "special" IR systems, thus, they employ Inverted Indexes too.

Distributed Indexing



These are bits that identify documents (Postings)

These are the terms (Lexicon)

Term Partition

Document Partition

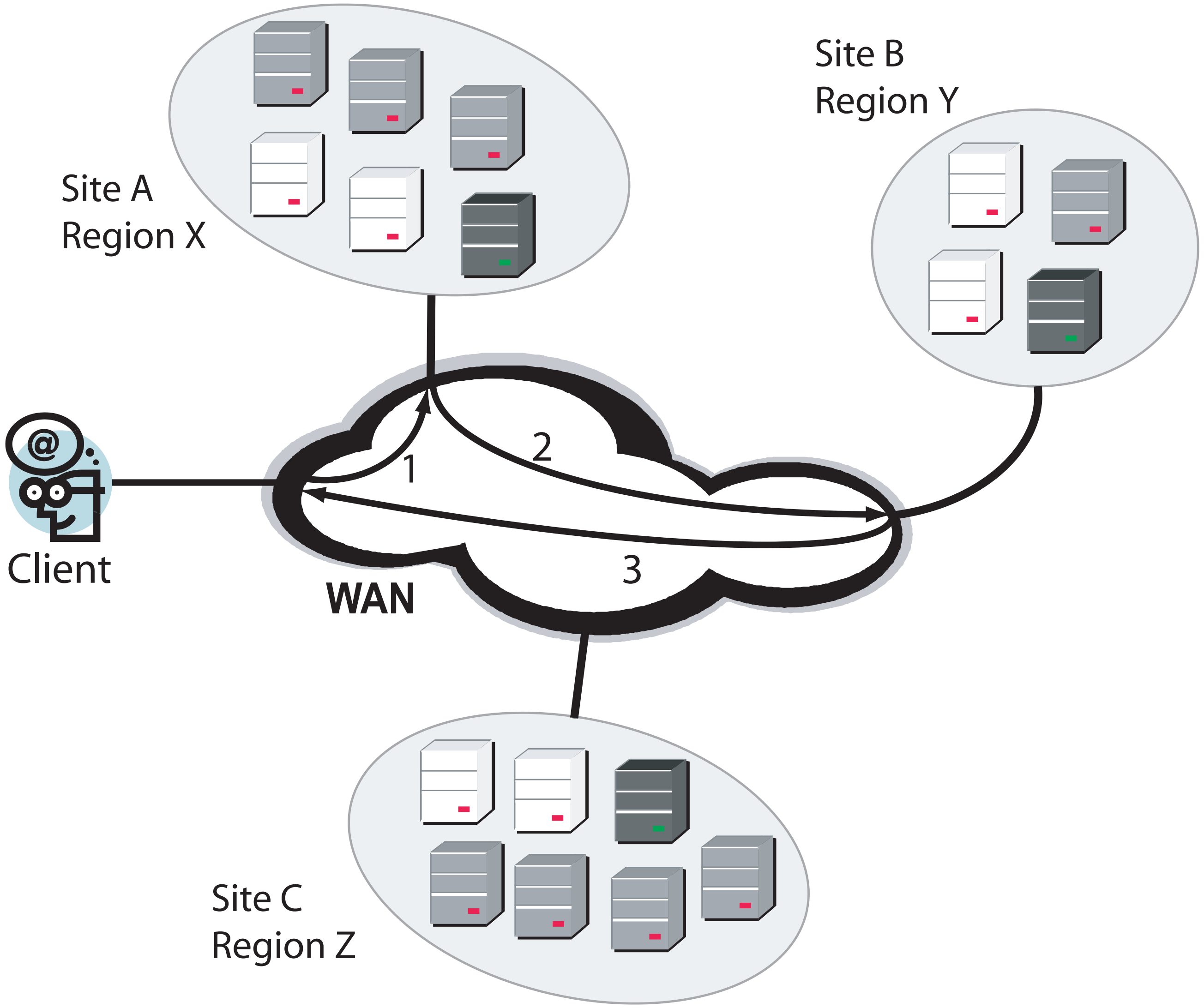
Query Processing

- System Components:
    - Clients submitting queries
    - Sites consisting of servers
    - Servers are commodity PCs
- Query Processing:
    - System receives a query
    - Query routing: forwarding queries to "*relevant*" sites
    - Merging results
- Challenges:
    - Determining appropriate sites on the fly
    - WAN communication is costly

Challenges in Details

- Large Scale Systems:
  - Large amount of data
  - Large data structures
  - Large number of clients and servers
- Partitioning of Data Structures:
  - Necessary due to very large data structure
  - Parallel processing
  - e.g. document collection split by topic, language, region
- Replication of Data Structures:
  - For availability, throughput and response time
  - Conflict with resource utilization

Framework for Distributed Query Processing



- Query Processor. Matches documents to received queries
- Coordinator. Receives queries and routes them to appropriate sites
- Cache. Stores results from previous queries

Currently...

- Multiple sites:
  - Sites are fully replicas of each other
  - Simple query routing: Dynamic DNS
- According to the previous framework:
  - Use storage resources more efficiently
  - More sophisticated query routing mechanisms
  - Effective partition strategies (e.g. language-based strategies)

For the Interested Readers

R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, F. Silvestri: Challenges on Distributed Web Retrieval. ICDE 2007: 6-20

The Lesson is Over