Data Mining

Keyword Based Search Engines (Part A)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

NJIT

New Jersey's Science & Technology University

THE EDGE IN KNOWLEDGE

DBMS, IR and DM

DBMS - search based on SQL attribute-value comparison

IR - search by topic or by keywords (recall, precision)

DM - extraction of knowledge from data

Search is essential to DM:

Find relevant data first and then mine the data to extract knowledge.

Difference between IR and DM:

- IR only searches data and returns results.

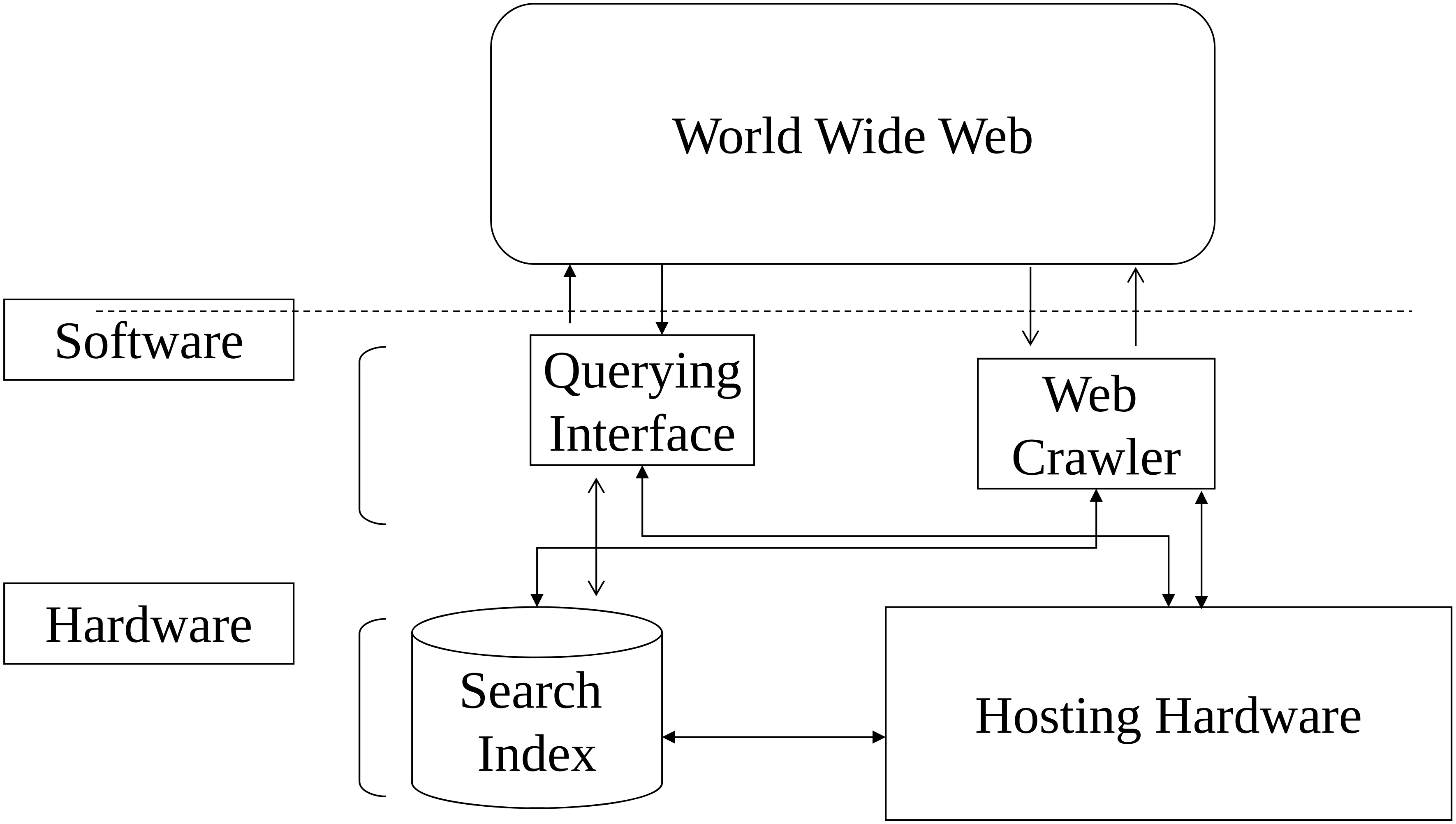
- DM discovers knowledge (e.g. a rule or association) from data.

Web Search Engines

Web search engines (or Web indexes, Index servers, search engines)

Popular ones: AltaVista, Excite, HotBot, Infoseek, Lycos, Yahoo!, Bing, Google

Architecture of Search Engine



Architecture of Search Engine (cont.)

- Querying Interface
  - Field Search
  - Keyword or Multiple-Term Query
  - Context Based Query
  - Natural Language Query

- Search Index

- Web Crawler

Field Search

For example, use title:information to search.

The following shows commonly used search fields.

Sno	Field Location	Example
Text	Body	Text: Information
Title:	Title	Title: Database
Link:	Hyperlink	Link: Kluwer.nl
Anchor:	Visual part of Hyperlink	Anchor: Mining
Url:	url	Url: www.xyz.com
Host:	Computer Name	Host: xyz.com
Domain:	Specific domain	Domain: Edu
Image:	Image Name	Image: Map.gif
Applet:	Applet Name	Applet: Tetris
Object:	Object Name	Object: game

Keyword or Multiple-Term Query:

E.g. database
E.g. information and database
E.g. ((data or Web) and mining)
E.g. Web or mining

Note: The order among words is unimportant.

Context Based Query:

Phrase search - "I have a dream"

Note: The order among words is important.

Proximity Querying:

AltaVista uses the near operator for two terms (they must be within 10 words of each other). This is useful in searching for names (first and last names that are separated by middle names and initials).

Natural Language Query

Ask Jeeves and ElectricMonk:
"Where can I find a digital camera?"
"Which models of cars are most popular?"
Pre-store millions of questions in knowledge bases.
Pre-built answers including lists of subsequent questions for narrowing the search.
If a search fails, control goes to its meta-search engine to retrieve various search engine results as a backup.

End of Keyword Based Search Engines Module (Part A)