Information Retrieval

Lecturer: CHEN Qingcai (qingcai.chen@hit.edu.cn)
TA: 汪安琪 田嘉豪 牛萌 杨鲁锋 张小童

Intelligent Computing Research Center,
Harbin Institute of Technology （Shenzhen）

"I believe there are reasons to think a Grand Theory of IR to be an unattainable goal - such a theory would have to encompass so many different aspects of retrieval, having to do for example with human cognition and behavior and the structure of knowledge, as well as with the statistical concepts that inform the probabilistic approach."

有足够的理由使我相信去考虑（建立）大统一的信息检索理论是一个难以企及的目标—这样的理论需要囊括信息检索的各个不同方面，比如需要包括人类认知学、行为学、知识结构，以及用在概率方法中的统计学概念等等。

--Stephen Robertson

on 2000's Salton Award lecture

Lecture 1
Introduction to Information Retrieval

References:

James Allan,University of Massachusetts Amherst
Amit Singhal, Google(R)

Outline

- What is Information Retrieval?
- Core idea of IR-related work

Questions before the course

- Which types of data had you processed?
- What is a database management system? Had you learnt at least one DBMS?
- Can traditional DBMS system manage text well? Why?

What is Information Retrieval?

Up to 2020-06-27 $408.8 Billion of Mkt Cap

Up to 2020-06-27 $41.98 Billion of Mkt Cap

- Unstructured data
- Quite effective (at some things)
- Highly visible (mostly)
- Commercially successful (some of them, so far)
- But what goes on behind the scenes?
  - How do they work?
  - Is there more to it than the Web?

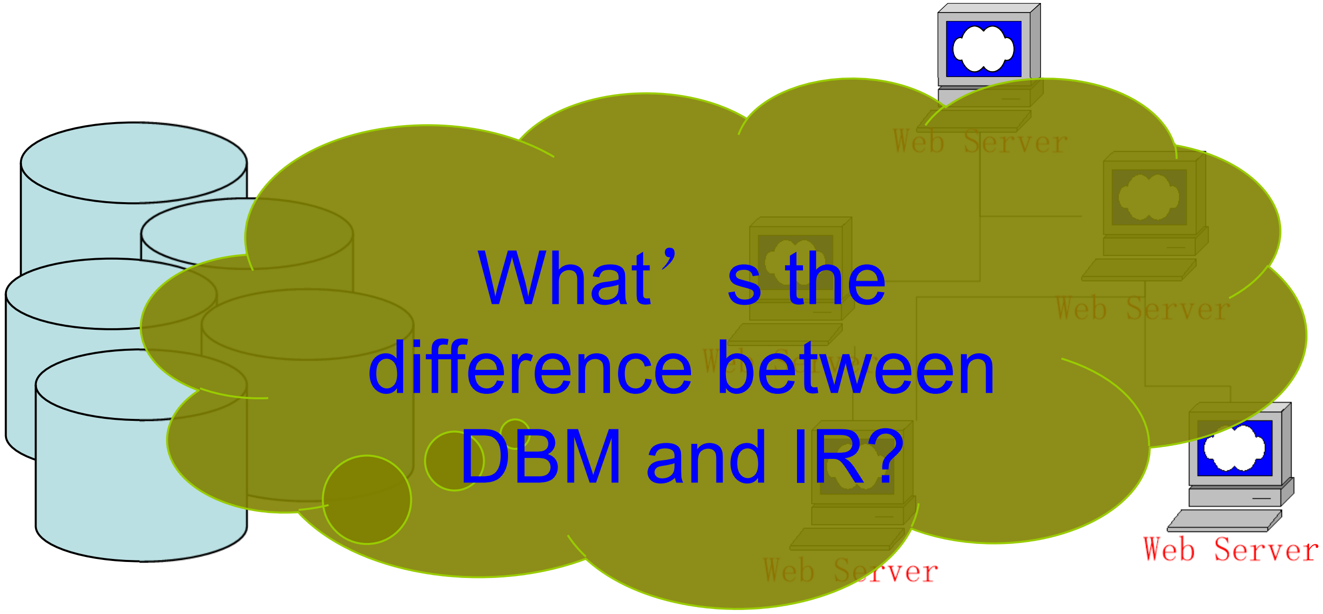
In this course, we ask…

- What makes a system like Google or BING tick?
  - How does it gather information?
  - What tricks does it use?
  - Extending beyond the Web
- How can those approaches be made better?
  - Natural language understanding?
  - User interactions?
- What can we do to make things work quickly?
  - Faster computers? Caching?
  - Compression?
- How do we decide whether it works well?
  - For all queries? For special types of queries?
  - On every collection of information?
- What else can we do with the same approach?
  - Other media?
  - Other tasks?

Tasks beyond document retrieval

- Question answering
- Agents (filtering, routing)
- Recommender systems
- Automatic organization (e.g., clustering)
- Cross-language retrieval
- Leveraging XML and other Metadata
- Data and information mining
- Knowledge management
- Meta-search (multi-database searching)
- Summarization
- …

Comparing IR to databases

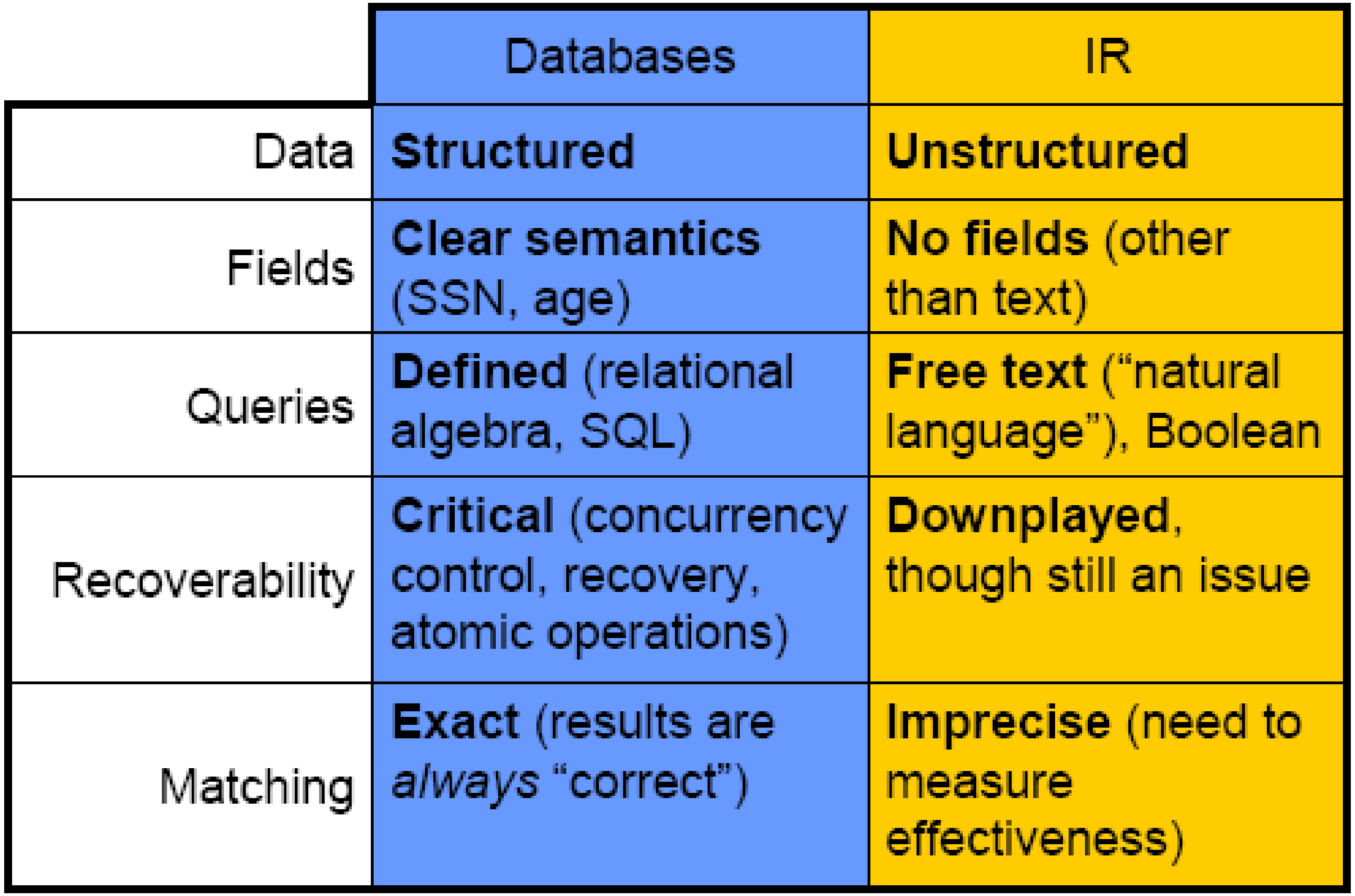


What’ s the difference between DBM and IR?

Database/Data warehouse

Information Retrieval

Comparing IR to databases



	Databases	IR
Data	**Structured**	**Unstructured**
Fields	**Clear semantics** (SSN, age)	**No fields** (other than text)
Queries	**Defined** (relational algebra, SQL)	**Free text** ("natural language"), Boolean
Recoverability	**Critical** (concurrency control, recovery, atomic operations)	**Downplayed**, though still an issue
Matching	**Exact** (results are *always* "correct")	**Imprecise** (need to measure effectiveness)

IR System

An unstructured data management system

Sample Systems

- IR systems
  - Inquery, Smart, Okapi, Lemur, Indri
- Database systems
  - Oracle, Informix, Access
- Web search and In-house systems
  - West, LEXIS/NEXIS, Dialog
  - Baidu, Bing, Sogou, AltaVista, Excite, Yahoo, Google, Dianping
- HotBot, Direct Hit, …
  - Ask Jeeves
  - eLibrary, GOV.Research_center, Inquira
- And countless others...

Outline

- *What is Information Retrieval?*
- Core idea of IR-related work

Basic Approach to IR

- Most successful approaches are statistical
  - Directly, or an effort to capture and use probabilities
- Why not natural language understanding?
  - i.e., computer understands documents and query and matches them
  - State of the art is brittle in unrestricted domains
  - Can be highly successful in predictable settings
    - Medical or legal settings with restricted vocabulary
- Could use manually assigned headings
  - e.g., Library of Congress headings, Dewey Decimal headings
  - Human agreement is not good
  - Hard to predict what headings are “interesting”
  - Expensive

Relevant(相关) Items are Similar(相似)

- Much of IR depends upon idea that
  - similar vocabulary → relevant to same queries
- Usually look for documents matching query words
- “Similar” (相似性) can be measured in many ways
  - String matching/comparison
  - Same vocabulary (词汇) used (?)
  - Probability that documents arise from same model
  - Same meaning of text

“Bag of Words”（词袋)

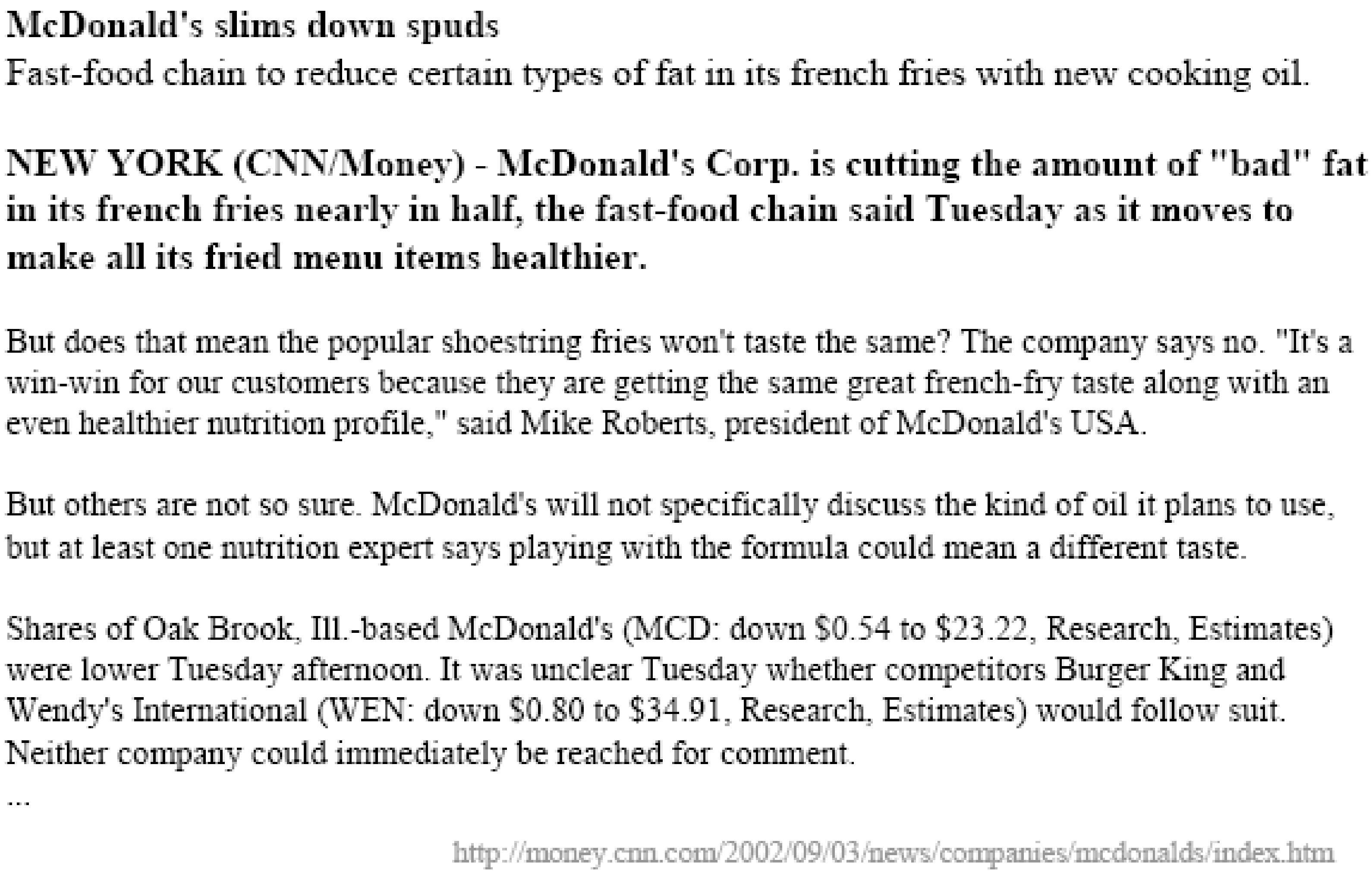
- An effective and popular approach
- Compares words without regard to order
- Example 1: consider reordering words in a headline……
  - **Random**: beating takes points falling another Dow 355
  - **Alphabetical**: 355 another beating Dow falling points
  - “**Interesting**”: Dow points beating falling 355 another
  - **Actual**: **Dow takes another beating, falling 355 points**

Example 2: What is this about?

16 × said 14 × McDonalds
12 × fat 11 × fries
8 × new 6 × company french nutrition
5 × food oil percent reduce taste Tuesday
4 × amount change health Henstenburg make obesity
3 × acids consumer fatty polyunsaturated US
2 × amounts artery Beemer cholesterol clogging director down eat estimates expert fast formula impact initiative moderate plans restaurant saturated trans win

1 × …
added addition adults advocate affect afternoon age Americans Asia battling beef bet brand Britt Brook Browns calorie center chain chemically … crispy customers cut … vegetable weapon weeks Wendys Wootan worldwide years York

Example 2(Cont’ d)
The original text (the start)



McDonald's slims down spuds

Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.

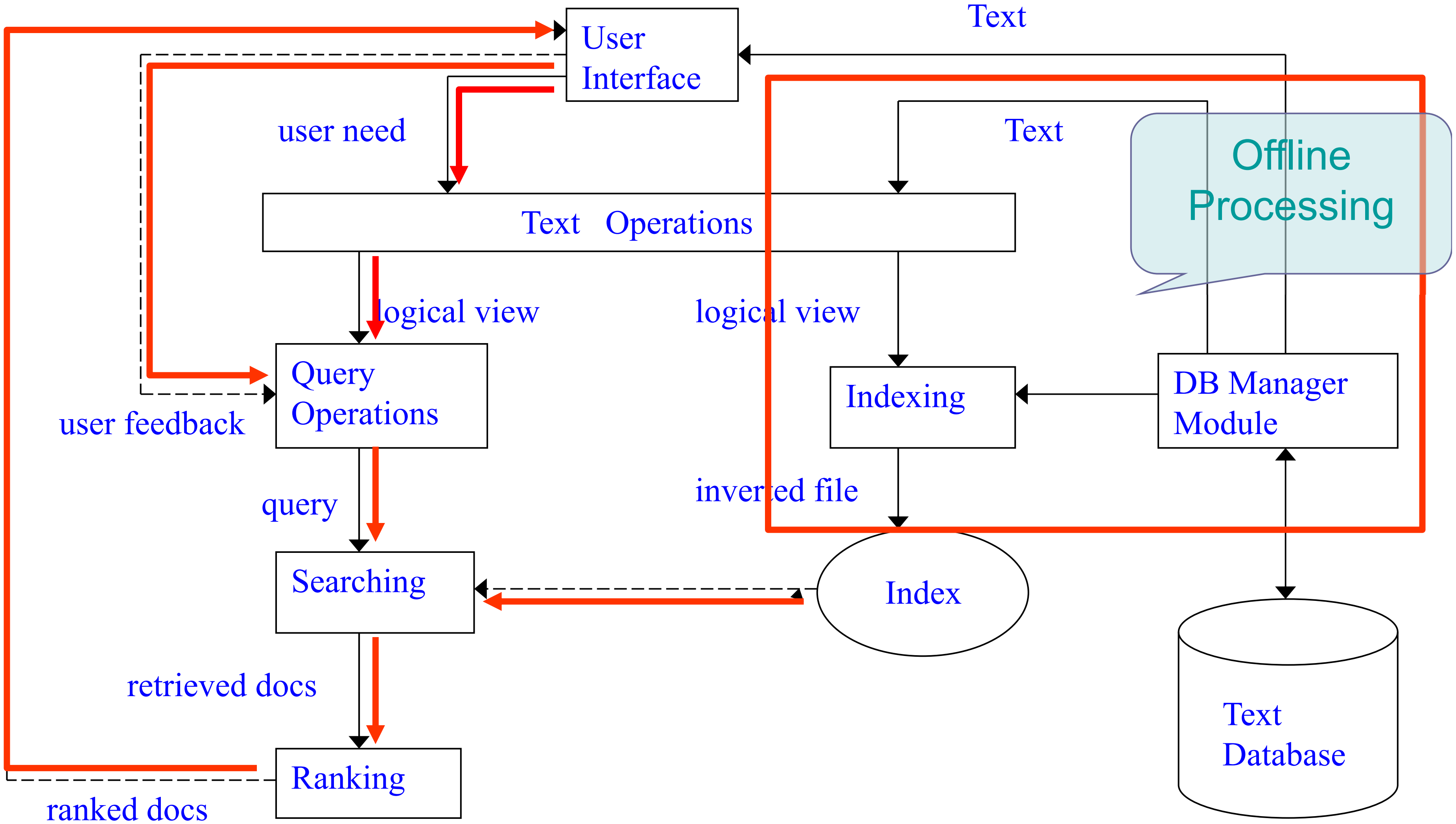
...

http://money.cnn.com/2002/09/03/news/companies/mcdonalds/index.htm

The Point from Example 2

- Basis of most IR is a very simple approach
  - find words in documents
  - compare them to words in a query
  - this approach is very effective!
- Other types of features are often used
  - phrases
  - named entities (people, locations, organizations)
  - special features (chemical names, product names)
    - difficult to do in general, especially in Chinese; usually require hand building
- Focus of research is on improving accuracy, speed
- …and on extending ideas elsewhere

Simple flow of retrieval process



Conclusion

- Information Retrieval?
  - Indexing, retrieving, and organizing text by probabilistic or statistical
- techniques that reflect semantics without actually understanding
  - Search engines
- Core idea
  - Bag of words captures much of the “meaning”
  - Objects that use vocabulary the same way are related

Baidu Brain