Introduction to Text Mining

Ali Ridho Barakbah

Knowledge Engineering Research Group
Department of Information and Computer Engineering
Politeknik Elektronika Negeri Surabaya

Definisi

- Menambang data yang berupa teks
- Sumber data biasanya didapatkan dari dokumen
- Tujuannya adalah mencari kata-kata yang dapat mewakili apa yang ada di dalam dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen

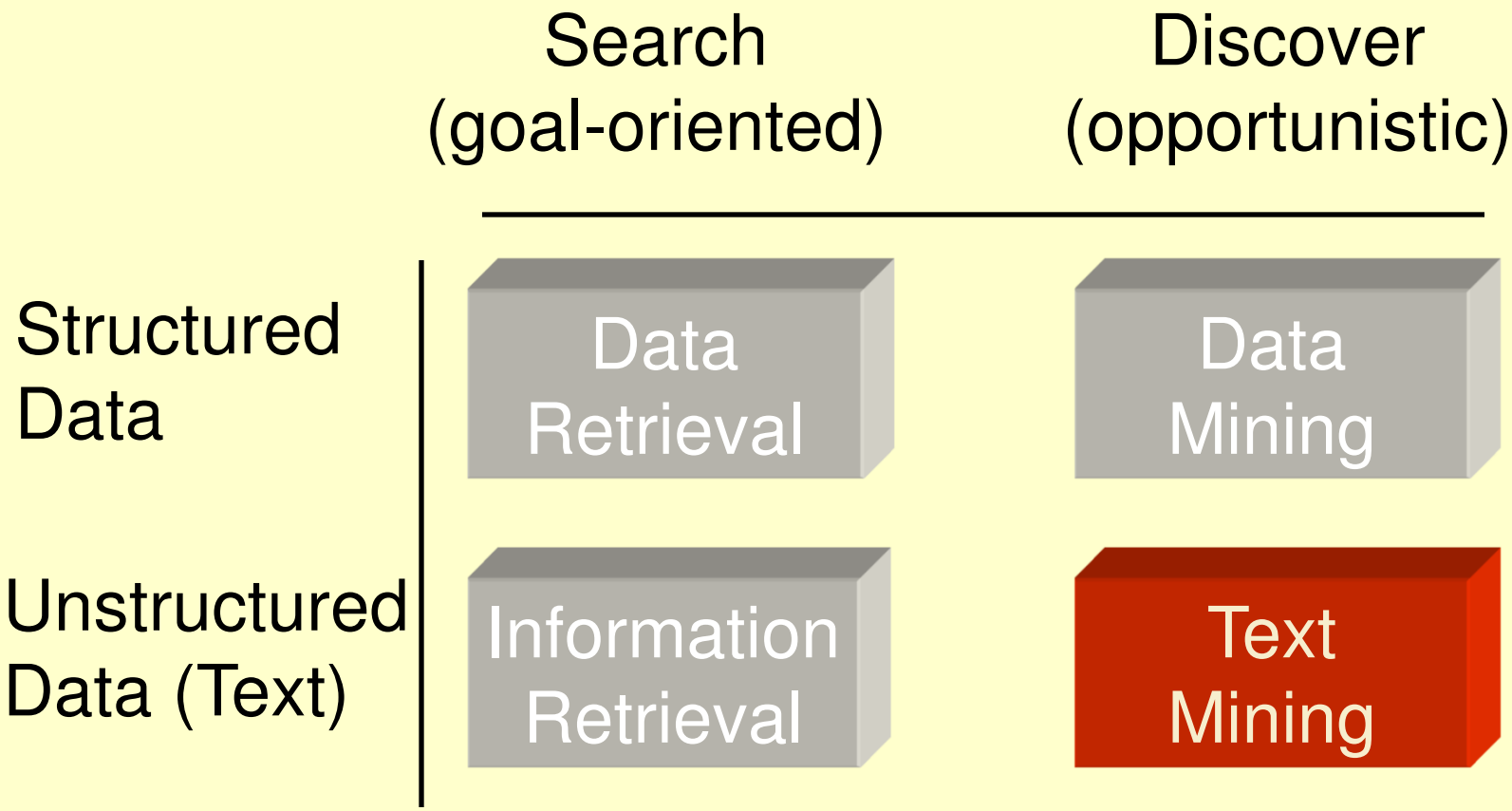
Keterkaitan Text Mining?

- Keterkaitan dengan data mining?
- Keterkaitan dengan computational linguistics?
- Keterkaitan dengan information retrieval?

	Finding Patterns	Finding "Nuggets"	
		Novel	Non-Novel
Non-textual data	General data-mining	Exploratory Data Analysis	Database queries
Textual data	Computational Linguistics		Information Retrieval

Source: Rebecca Hwa, Overview of Text Mining, 2002

"Search" versus "Discover"



Data Retrieval

- Find records within a structured database.

Database Type	Structured
Search Mode	Goal-driven
Atomic entity	Data Record
Example Information Need	"Find a Japanese restaurant in Boston that serves vegetarian food."
Example Query	"SELECT * FROM restaurants WHERE city = boston AND type = japanese AND has_veg = true"

Information Retrieval

- Find relevant information in an unstructured information source (usually text)

Database Type	Unstructured
Search Mode	Goal-driven
Atomic entity	Document
Example Information Need	"Find a Japanese restaurant in Boston that serves vegetarian food."
Example Query	"Japanese restaurant Boston" or Boston->Restaurants->Japanese

Data Mining

- Discover new knowledge through analysis of data

Database Type	Structured
Search Mode	Opportunistic
Atomic entity	Numbers and Dimensions
Example Information Need	"Show trend over time in # of visits to Japanese restaurants in Boston "
Example Query	"SELECT SUM(visits) FROM restaurants WHERE city = boston AND type = japanese ORDER BY date"

Text Mining

- Discover new knowledge through analysis of text

Database Type	Unstructured
Search Mode	Opportunistic
Atomic entity	Language feature or concept
Example Information Need	"Find the types of food poisoning most often associated with Japanese restaurants"
Example Query	Rank **diseases** found associated with "Japanese restaurants"

Challenges of Text Mining

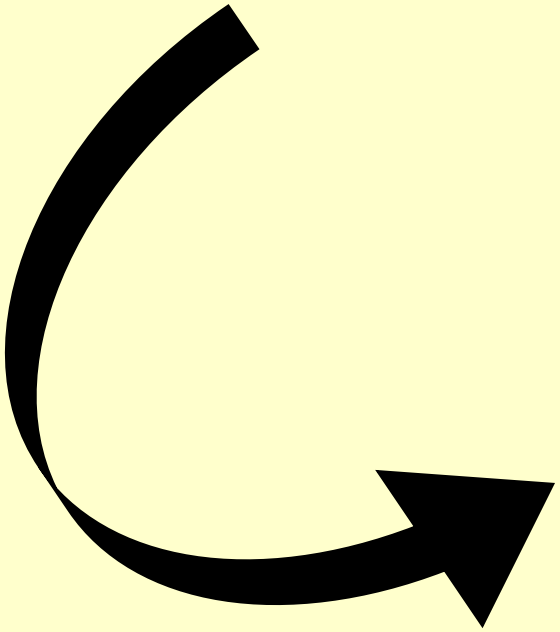
- Very high number of possible “dimensions”
  - All possible word and phrase types in the language!!
- Unlike data mining:
  - records (= docs) are not structurally identical
  - records are not statistically independent
- Complex and subtle relationships between concepts in text
  - “AOL merges with Time-Warner”
  - “Time-Warner is bought by AOL”
- Ambiguity and context sensitivity
  - automobile = car = vehicle = Toyota
  - Apple (the company) or apple (the fruit)

Tahapan

- Tokenizing
- Filtering
- Stemming
- Tagging
- Analyzing

Tokenizing

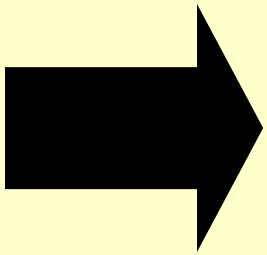
This lecture is talking about how to mine data



this
lecture
is
talking
about
how
to
mine
data

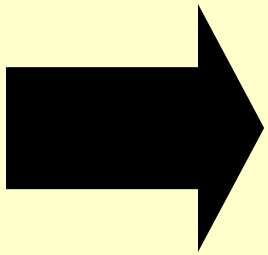
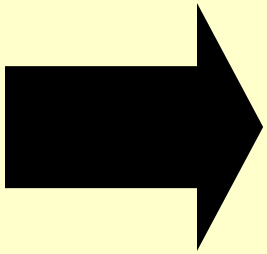
Filtering

this
lecture
is
talking
about
how
to
mine
data

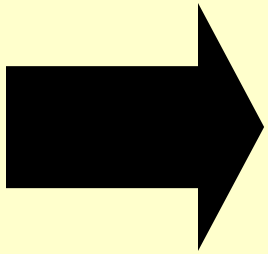


lecture
talking
mine
data

Stemming

lecture talking mine data		lecture talk mine data
reading stories		read stori

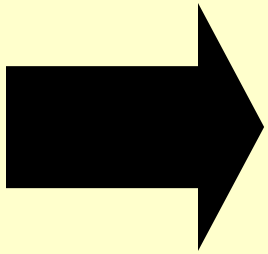
Tagging

thougt was stori		think be story

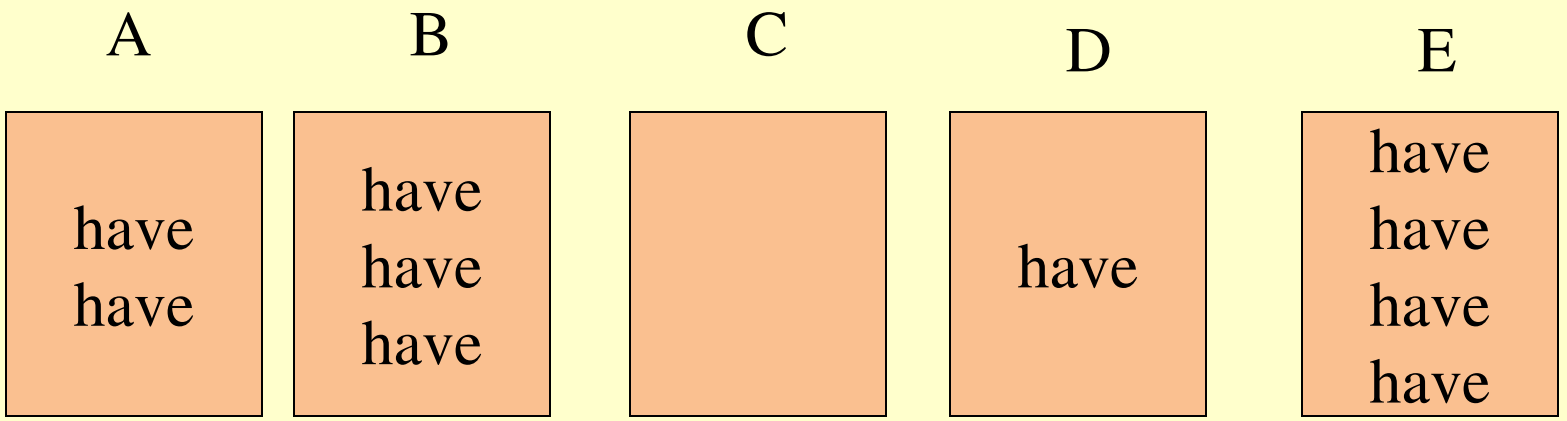
Analyzing

- Mencari seberapa jauh keterhubungan antar kata-kata antar dokumen
- Term Frequency-Inversed Document Frequency (TF-IDF) → Algoritma yang paling sederhana yang biasanya dipakai untuk scoring

lecture
talk
mine
data



Lecture → 0.8
Talk → 0.34
Mine → 0.7
Data → 0.45



$$TFIDF_{d,t} = FREQ_{d,t}(1 + \log \frac{N}{DFREQ_t})$$

$$\text{TFIDF}_{have,B} = 3 \text{ x } (1 + \log(5 / 4))$$

Result of Text Mining