Text Mining

Data Mining

Database vs. Information Retrieval

- Database Management Systems (DBMS)
  - Management of data
    - Maintain data to ensure data correctness
    - Provide query functionality
  - Structured data: fixed attributes
- Information Retrieval (IR)
  - Management of text, document
  - Un-structured data: free text

An Example: 學生修課資料庫

Student

? ?	? ?	? ?	? ?	? ?
1101	徐懷鈺	女	1978/3/3	yuki
2301	孫燕姿	女	1978/7/23	美少男殺手的對手
1102	卜學亮	男	1969/9/11	凡走過必留下痕跡
1201	蔡依林	女	1980/9/15	美少男殺手
1103	劉若英	女	1973/6/1	奶茶
1301	金城武	男	1973/10/11	美少女殺手
2302	周杰倫	男	1979/1/18	新美少女殺手

SC

? ?	? ?	? ?
1101	C3001	90
1102	C3001	70
1102	J2010	80
1103	C3001	100
2301	J2025	85
2301	C3001	90
2302	J2010	70
2302	J2025	80
1301	C3001	80
1301	J2010	85

Course

? ?	? ?	? ?
C3001	? ? ? ? ? ?	3
J2010	? ? ? ? ? ?	4
C3020	? ? ? ? ?	2
J2025	? ? ? ? ? ?	3

An Example: Text

Wii是任天堂公司的家用遊戲主機。Wii是Game Cube的後繼機種，屬於第七世代家用遊戲機，同時期的競爭對手是微軟的Xbox 360及Sony的PlayStation 3。Wii是任天堂所推出的第五部家用遊戲機（前四部為紅白機、超級任天堂、任天堂64、GameCube），其主要特色為前所未見的控制器使用方法、懷舊遊戲主機軟體下載販賣及待機時網路連線等。

...

Information Retrieval

- Text retrieval
- Text (document)
  - Unstructured data, not structured data
  - Text retrieval
    - Retrieval by text content
      - e.g. retrieve the documents which contain the words "database" or 'multimedia"
    - Ranking, relevant retrieval, not exact matching only
      - e.g. retrieve the books which related to "database" and "multimedia"
  - Text browsing

Information Retrieval vs. Information Filtering

- Information Retrieval
  - Ad hoc search
  - The documents in the collection remain relatively static while new queries are submitted to the system
  - Pull
- Information filtering
  - Routing, Recommendation
  - The queries remain relatively static while new documents come into the system
  - User profile is kept to filter information
  - Push

Mathematical Models for IR

- Mathematical Models for Information Retrieval
  - Set Theoretic Models
    - Boolean model
    - Fuzzy-set model
    - Extended Boolean model
  - Algebraic Model
    - Vector space model
    - Generalized vector space model
    - Latent Semantic Index model
    - Neural networks
  - Probabilistic Model
  - Hybrid model

Boolean Model

Boolean Model

- Document
  - is modeled as a set of index terms (Boolean variable)
  - Bag of Words Model
  - e.g. D_1= {data, structure, video}

    D_2={multimedia, data, audio, VRML}
- Query
  - is modeled as a Boolean expression of query terms

  <e.g.> Q = (data AND structure) OR

  (multimedia AND NOT video)
  - each document is either relevant or non-relevant to the query

Boolean Model (Cont.)

- Advantage: simple to implement
- Disadvantage
  - – exact matching, no ranking result
  - – may lead to retrieval of too few or too many documents
  - – no weight assignment to query terms
  - – Relevance feedback?
  - * Relevance feedback
    - automatic query refinement
    - derived from user’s feedback on system generated results

Vector Space Model

Vector Space Model

- Document
  - each index term is associated with a positive weight
  - each document is represented as a vector of indexed terms

$$D_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,t})$$

where t = total no. of index term s in the system

<e.g.> (data, structure, video, Python, audio, MPEG)

$D_1 = (6 , 4 , 10 , 0 , 0 , 2)$

$D_2 = (4 , 0 , 1 , 10 , 8 , 6)$

- Query: a vector of query terms

<e.g.>

$Q = (10 , 8 , 0 , 8 , 0 , 0)$

Vector Space Model (cont.)

- Similarity measure: degree of similarity

$$Q \bullet D_i \Big/ |Q||D_i|$$

<e.g.> (data, structure, video, Python, audio, MPEG)

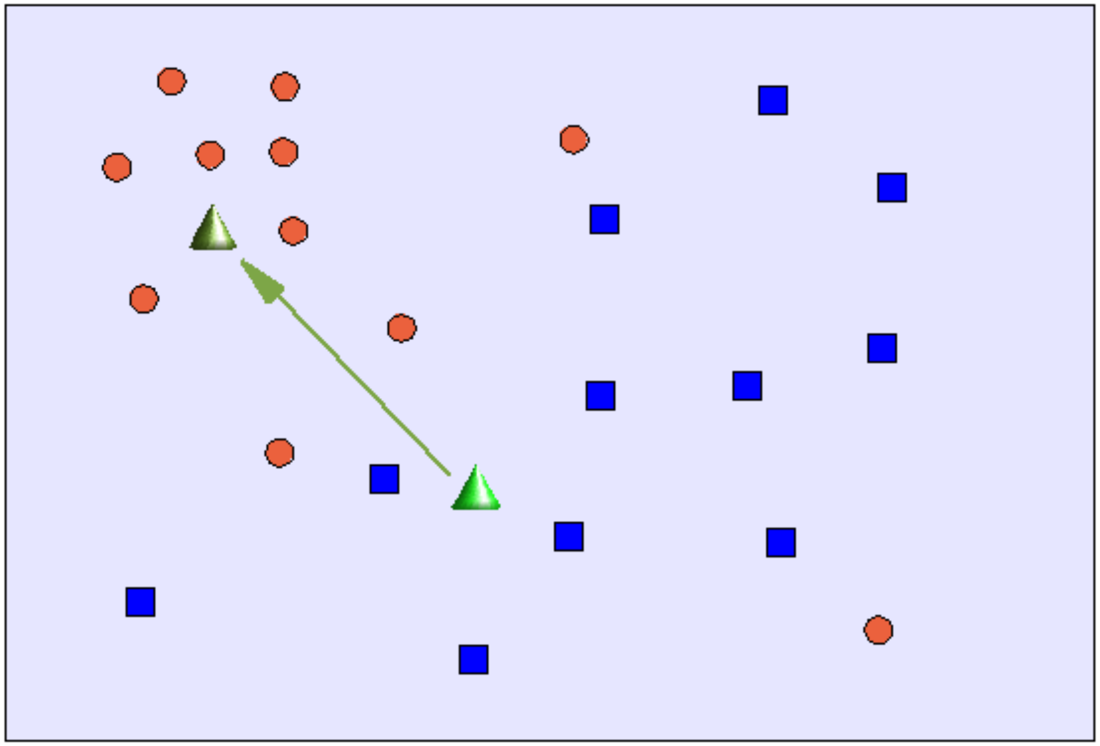
$D_1 = ($ 6 , 4 , 10 , 0 , 0 , 2 $)$

$Q = ($ 10 , 8 , 0 , 8 , 0 , 0 $)$

$$Q \bullet D_1 = \frac{\mathbf{10*6+8*4+0*10+8*0+0*0+0*2}}{\sqrt{\mathbf{10^2+8^2+0^2+8^2+0^2+0^2}}\sqrt{\mathbf{6^2+4^2+10^2+0^2+0^2+2^2}}}$$

Vector Space Model (cont.)

- Advantages
  - ranking.
  - relevance feedback (query refinement by vector modification)



Automatic Indexing

検索

Automatic Indexing

- Indexing: term extraction (key-word extraction)
- Indexing method
  - manual annotation
  - automatic indexing

    Step 1. Parsing (segmentation)

    Step 2. Stop-list removal (common words)

    Step 3. Stemming (suffix, prefix)

    Step 4. Phrase & Synonyms (Thesaurus)

    Step 5. Weight Judgment

Parsing

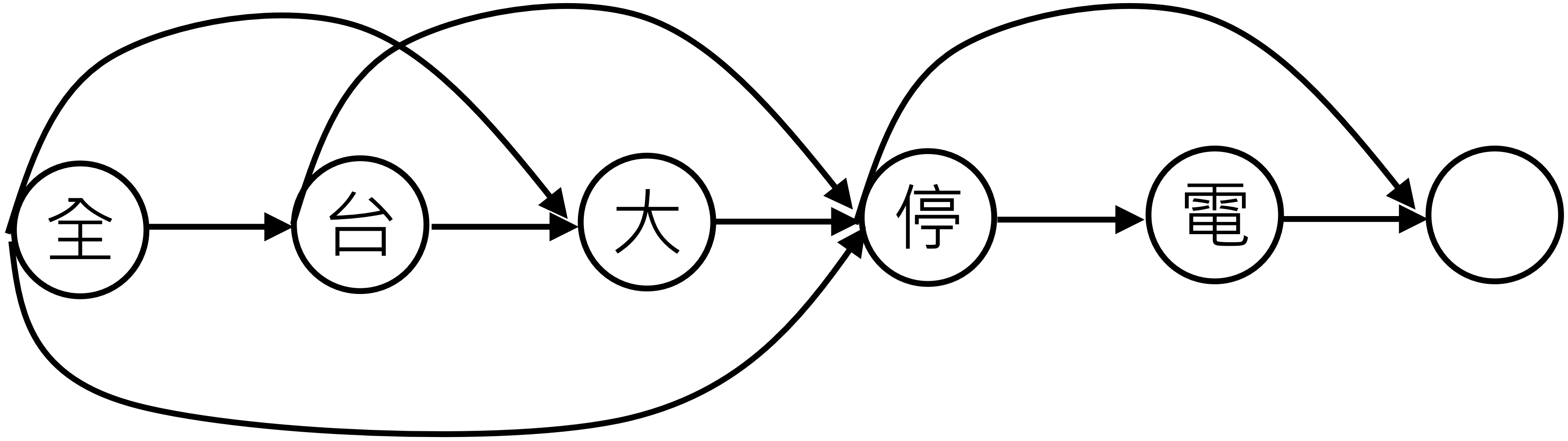
- Lexical analysis of text
- Segmentation (斷詞)
  - 這名記者會說國語
  - 這名 記者 會 說 國語
- Word separators
  - space
  - digits
  - hyphens
  - punctuation marks
  - the case of the letters

Parsing (cont.)

- Ambiguity of Segmentation
  - 民可使由之不可使知之
    1. 民可使由之，不可使知之。
    2. 民可，使由之；不可，使知之；
    3. 民可使，由之；不可使，知之。
    4. 民可使，由之不可，知之。
    5. 民可使由之？ 不。 可使知之。
    6. 民可使由之？ 不可。 使知之。
  - 全台大停電
  - 近日報載台大及中央大學接連出現校園竊賊，嫌犯偽裝成大學生進入實驗室
  - 到了103年國三的四月間，則要參加全國性教育會考
  - 文金會上將討論無核問題
  - 稱終身定期金契約者，謂當事人約定，一方於自己或他方或第三人生存期內，定期以金錢給付他方或第三人之契約。

Parsing (cont.)

- Approaches of Segmentation
  - Dictionary-based approach
    - 全台大停電



  - Statistics-based approach
  - Linguistic approach
  - Deep Learning approach

Stop-List Removal

- Stop-list
  - a list of stop words
  - words that are too frequent among the documents

    a the at on by and but
  - article, prepositions, conjunctions,….
- Can reduce the size of the indexing structure considerably
- Problem
  - Search for “to be or not to be”?

Stemming

- Example
  - *connect, connected, connecting, connection, connections*
  - effectiveness, effective, effect
  - picnicking, picnic
  - king, k ?
- Removing strategies
  - affix removal: intuitive, simple
  - table lookup
  - successor variety
  - n-gram

Indexing

$\left\{ \begin{array}{l} i : \text{term} \\ j : \text{document} \end{array} \right.$

- Key words: high discrimination
- Frequency-based indexing (TF-IDF)
  - TF (term frequency) tf_{ij}

    : frequency of term T_j in document D_i

    * normalized TF
  - IDF (Inverse-document frequency) df_j

    : document frequency of term T_j in a collection of N documents.
  - Weight of term T_j in document D_i

when $df_j \approx N$

$(\log 1 = 0)$

$$w_{ij} = \frac{tf_{ij}}{df_j} \qquad\qquad w_{ij} = \underset{tf}{\underline{tf_{ij}}}\ \underset{\text{idf (weight)}}{\underline{\log\left(\frac{N}{df_j}\right)}}$$

Text Mining

- Text Categorization
- Sentiment Analysis
- Opinion Mining
- Named Entity Recognition (NER)
- Disambiguation
- Text Summarization
- Event Detection
- Fake News Detection
- Natural Language Processing
- …