Text Categorization

Ko, Youngjoong

Sungkyunkwan University

Overview of this lecture

- Text Representation & Similarity Calculation
- The theory of text categorization

Warming up!!

- Pattern classification (Duda & Hart)

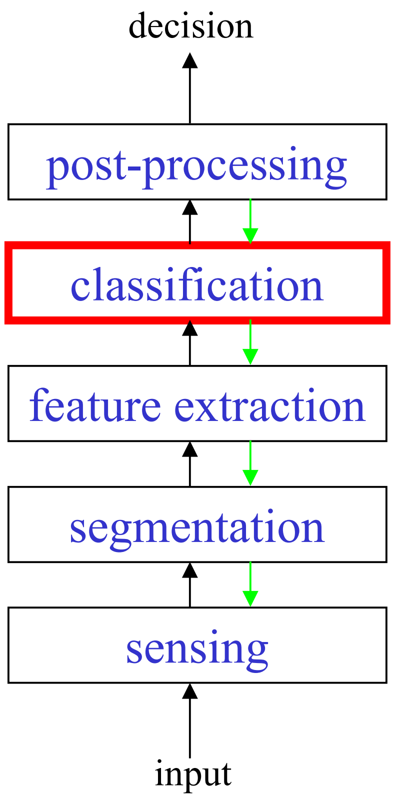


Fig1. The process of the pattern recognition system

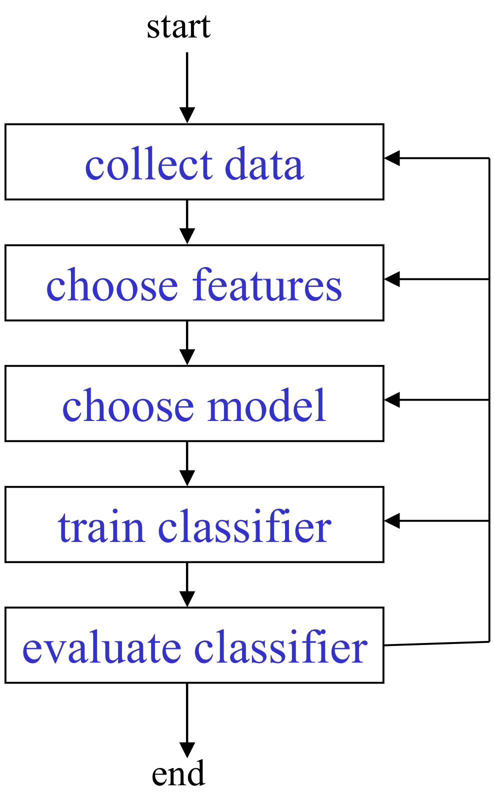
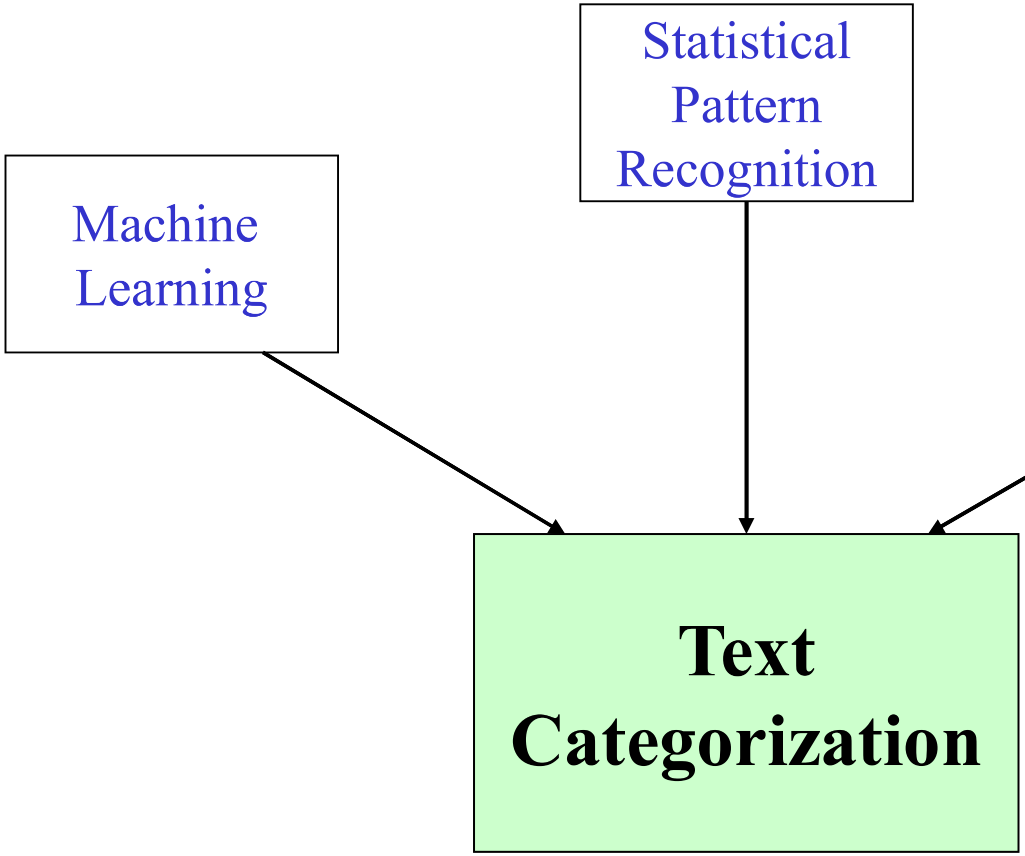


Fig2. The design cycle of the pattern recognition system

Warming up!!



Statistical Pattern Recognition

Machine Learning

Information Retrieval

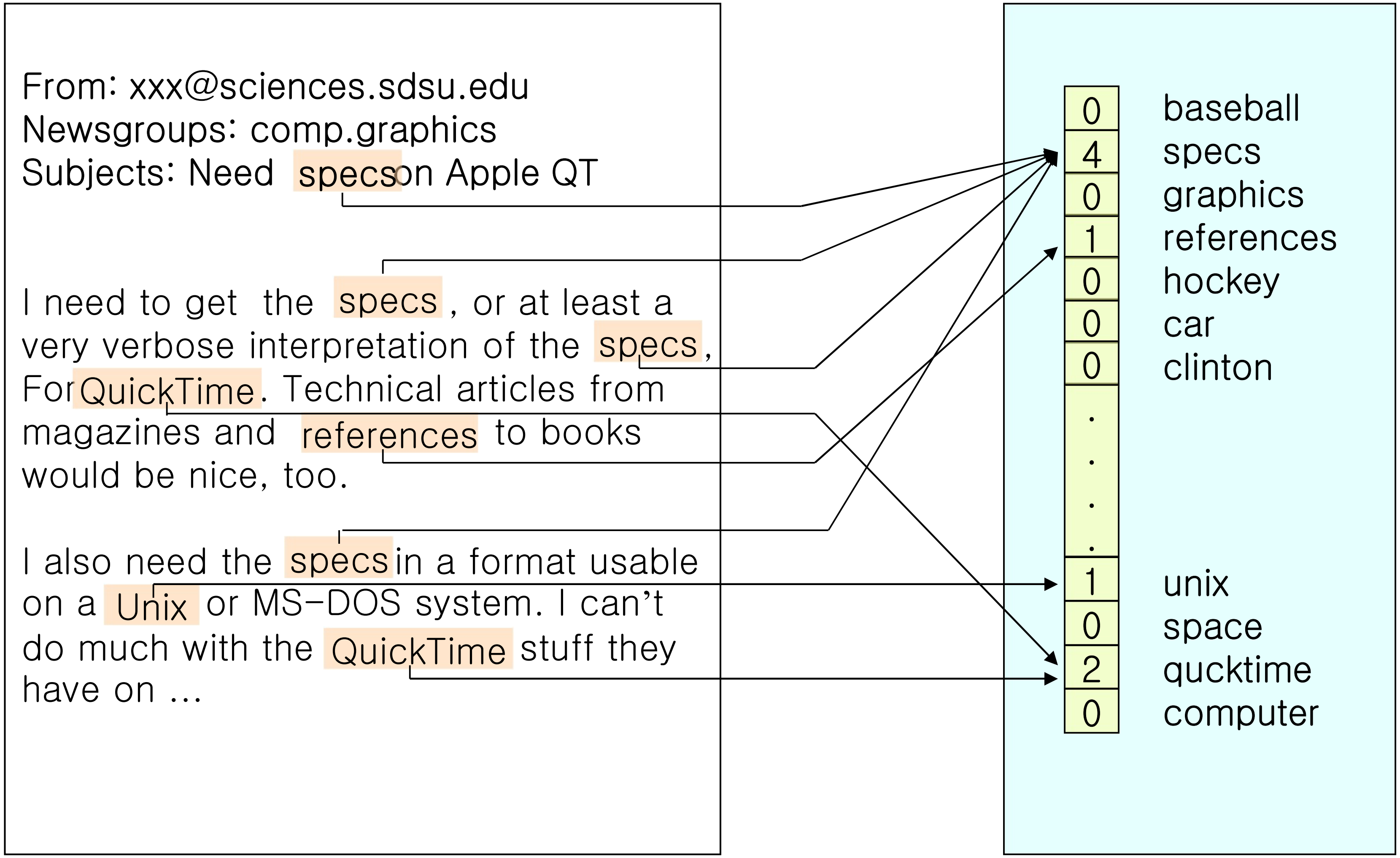
Text Categorization

Text Representation

Ko, Youngjoong

Sungkyunkwan University

Important Words and their Important Scores



From: xxx@sciences.sdsu.edu
Newsgroups: comp.graphics
Subjects: Need specs on Apple QT

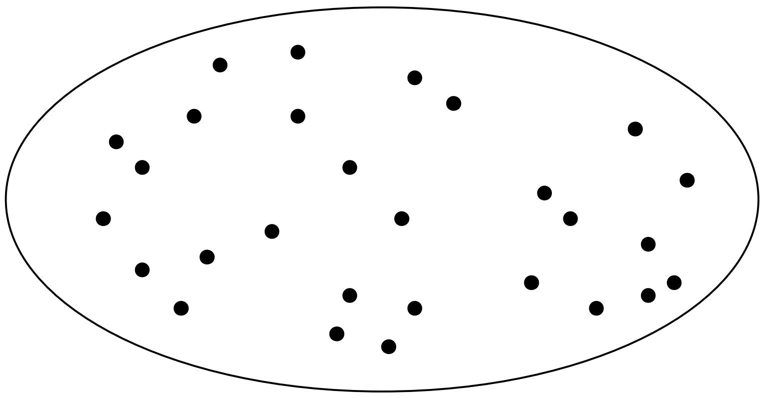
I need to get the specs, or at least a very verbose interpretation of the specs, For QuickTime. Technical articles from magazines and references to books would be nice, too.

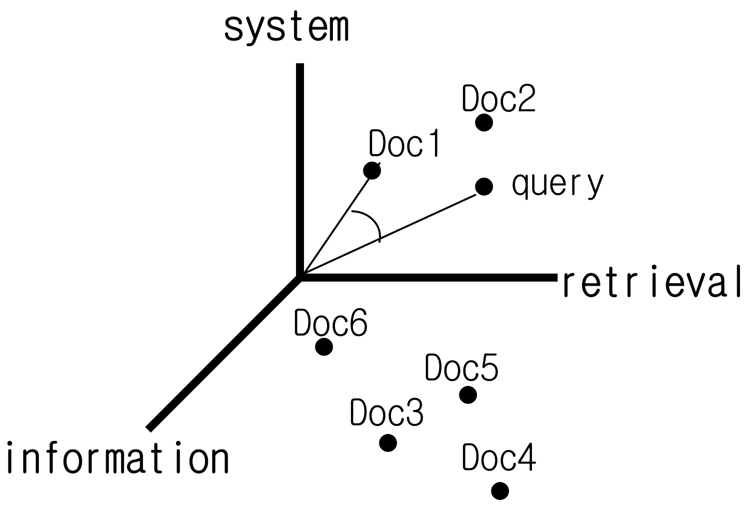
I also need the specs in a format usable on a Unix or MS-DOS system. I can't do much with the QuickTime stuff they have on …

0	baseball
4	specs
0	graphics
1	references
0	hockey
0	car
0	clinton
.	
.	
.	
.	
1	unix
0	space
2	qucktime
0	computer

Vector Space Model

- In the multi-dimensional space
  - To represent document as a vector
    - Become a document to a point in the vector space model
  - Each Dimension
    - Term or concept





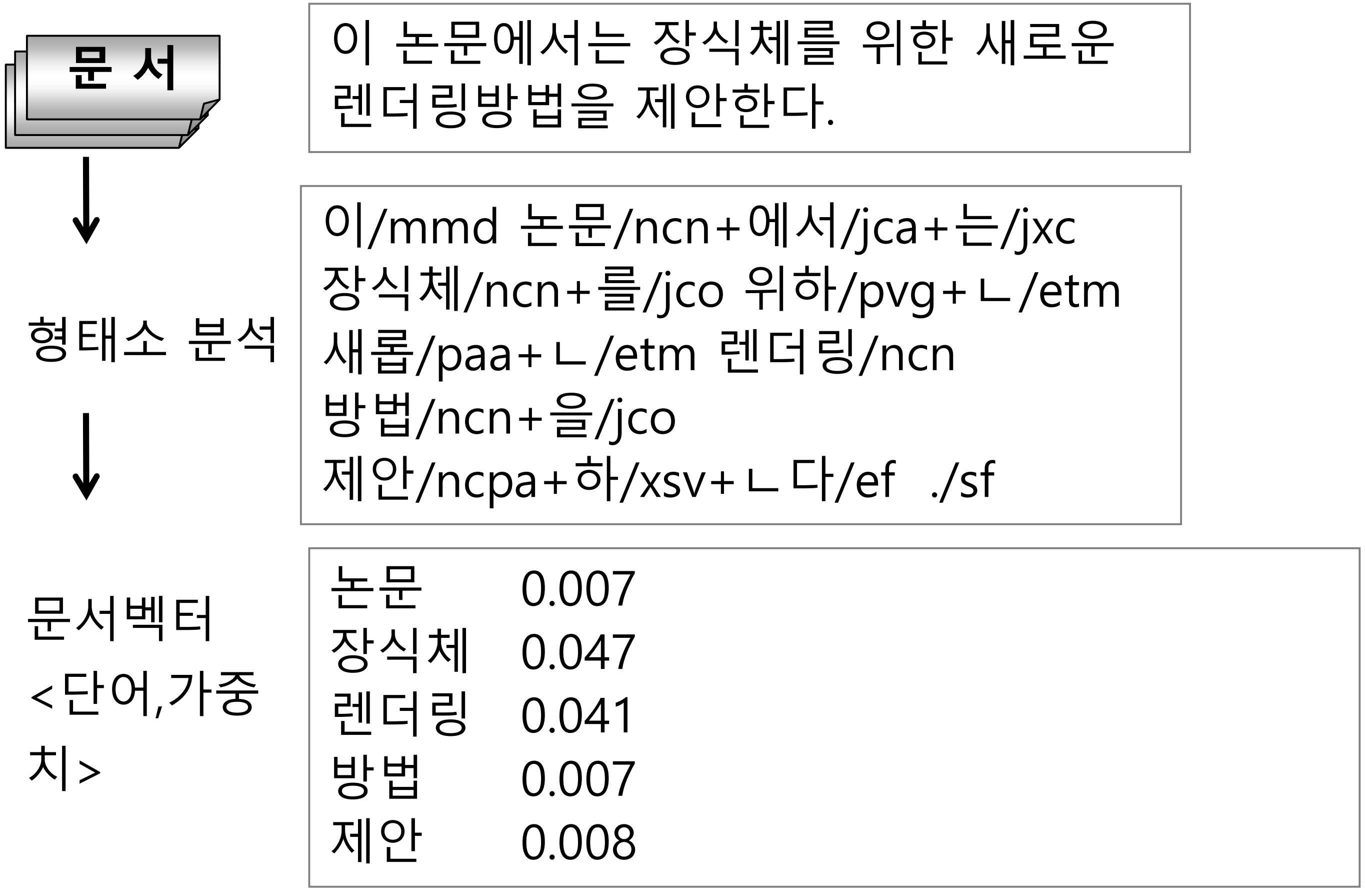
Feature Extraction

- Korean
  - POS tagging
  - Noun extraction
    - 학교/ncn + 에서/jca -> 학교
  - Removing stop words

- English
  - Removing stop words
    - a, the, this, …
  - stemming
    - swimming, swims, swimmer –> swim
    - flowers –> flower

Example for Vector Representation

- An example using POS tagging



문 서

이 논문에서는 장식체를 위한 새로운 렌더링방법을 제안한다.

형태소 분석

이/mmd 논문/ncn+에서/jca+는/jxc
장식체/ncn+를/jco 위하/pvg+ㄴ/etm
새롭/paa+ㄴ/etm 렌더링/ncn
방법/ncn+을/jco
제안/ncpa+하/xsv+ㄴ다/ef ./sf

문서벡터 <단어,가중치>

논문	0.007
장식체	0.047
렌더링	0.041
방법	0.007
제안	0.008

Term Weight Calculation

- Important factors of estimating TFIDF term weights
  - How many does the term occur in the document?
    - The more occurrence of a term appear in the document, the more importance of the term
    - Term Frequency (TF)
  - Is this common term or technical term?
    - Technical term is more important.
    - Inverted Document Frequency (IDF)
  - The length of the document
    - If a term occur two times in 10 words document OR 20 words document… which one has higher term weight?
    - Normalization of document length

Term Weighting Scheme

- Term weight calculation formulae
  - *Term Frequency (tf)*

    tf_t **Term frequency of term, *t, in a document***

  - Inverse Document Frequency (*idf*)

    $$idf_t = \log_2 \frac{N}{df_t}$$

    N **: total number of documents**

    df

- *Tfidf* term weighting formula

$$weight_t = tf_t \cdot idf_t$$

Term Weighting Scheme

- Normalization of document length

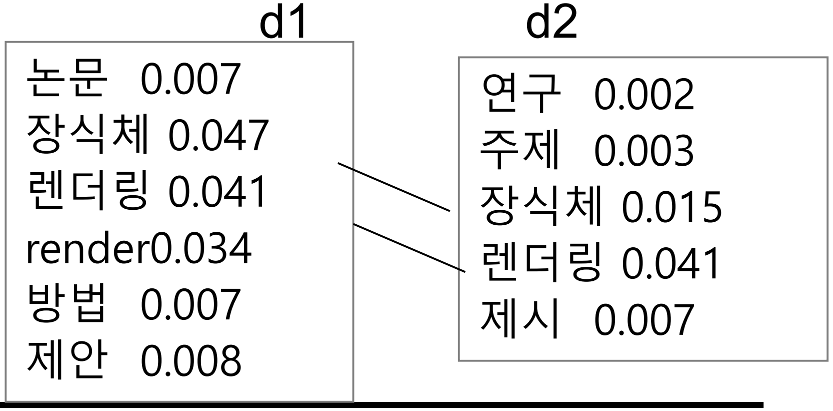
$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{r} (tfidf(t_s, d_j))^2}}$$

Real Document Representation

- Document Vector
  - N dimension: total number of terms in total corpus (N)
    - <weight>
    - Few terms in a document
    - Most terms have 0 weight.

	d1	d2
논문	0.007	0
연구	0	0.002
주제	0	0.003
장식체	0.047	0.015
렌더링	0.041	0.041
render	0.034	0
방법	0.007	0
제시	0	0.007
제안	0.008	0
…	0	0
…	0	0

- Document vector in real application
  - Using only appeared terms
    - <term, weight>
  - When similarity measurement
    - Search same words in both docs.



d1

논문	0.007
장식체	0.047
렌더링	0.041
render	0.034
방법	0.007
제안	0.008

d2

연구	0.002
주제	0.003
장식체	0.015
렌더링	0.041
제시	0.007

Similarity Measure Method

- Similarity Measures
  - Quantity that reflects the strength of relationship between two objects

- Similarity Measure Methods
  - Inner product
  - Euclidean distance
  - Cosine coefficient

Similarity Measure Method

- Inner product
  - The basic method between query and document in Information Retrieval

$$sim(d_i, d_j) = \sum_{k=1}^{n} w_{ik} \cdot w_{jk}$$

- Euclidean distance
  - The less distance value, the more similar

$$dist(d_i, d_j) = \sqrt{\sum_{k=1}^{n} (w_{ik} - w_{jk})^2}$$

Similarity Measure Method

- Cosine coefficient
  - Normalized inner product
    - Similarity value range : [0 ~ 1]
      - 1: two documents are same
      - 0: there is no co-occurred term between two documents

$$sim(d_i, d_j) = \frac{\sum_{k=1}^{n} w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \cdot \sum_{k=1}^{n} w_{jk}^2}}$$

d

d_i **벡터** $w_{i1}, w_{i2}, \cdots, w_{in}$

d_j **벡터** $w_{j1}, w_{j2}, \cdots, w_{jn}$