Data Mining & Text Mining

2010.10.28

동국대학교 통계학과
이영섭
yung@dongguk.edu

이 발표자료의 일부는 노현정연구원((주)사이람)과의 공동연구에 기초한 것입니다.

Data Mining

KDD

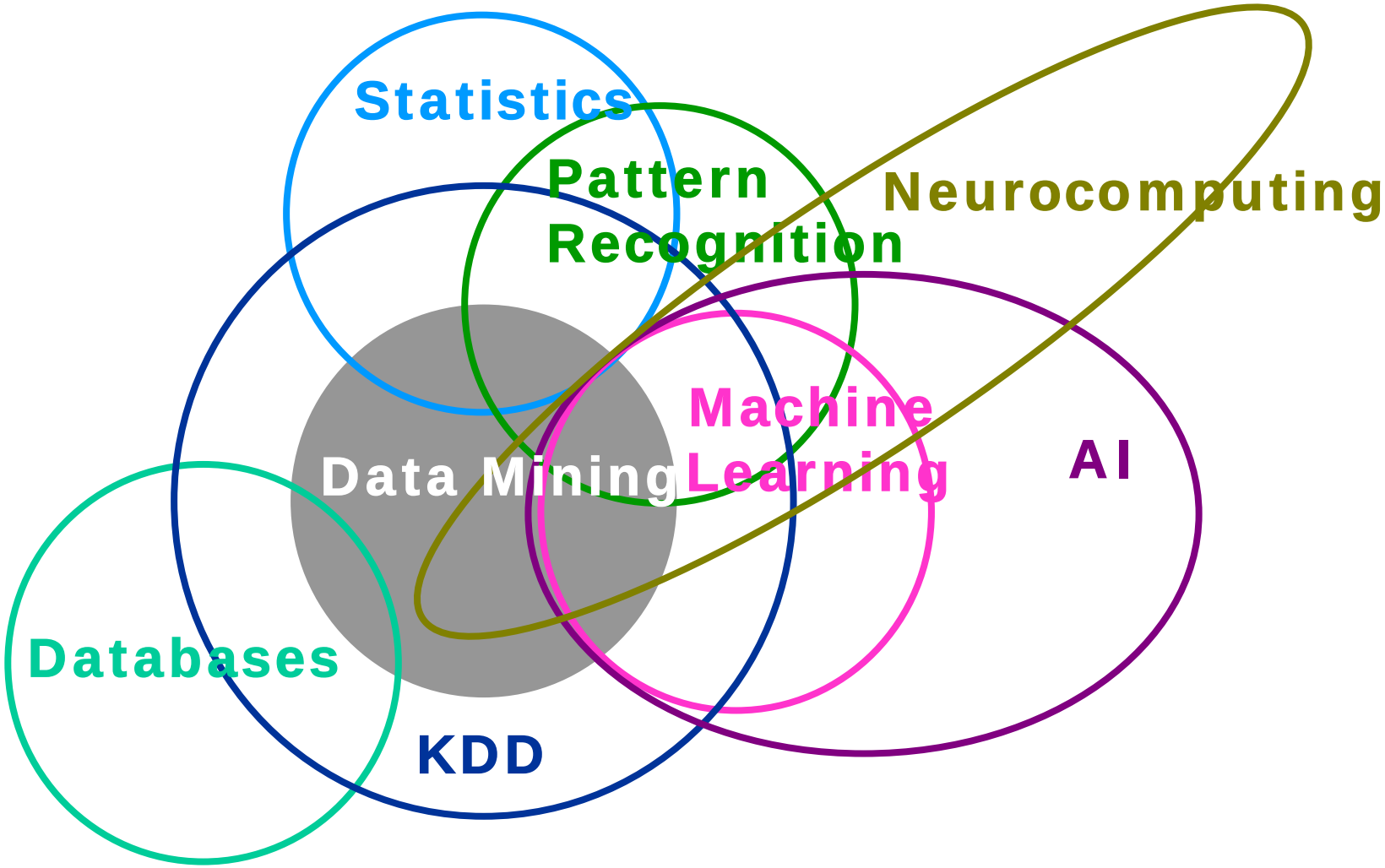
- Knowledge Discovery in Databases (KDD,1989)
  - Overall process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data
  - KDD conference on Knowledge Discovery and Data Mining (from 1995) Change the name to KDD conference on Knowledge Discovery in Data (from 1999) (rapidly growing conference- paper's accept rate 20%)
  - Multidisciplinary research area including Databases, Data Warehouse, Statistics, pattern recognition, machine learning (a branch of AI), information science and neurocomputing(neural networks)

What is Data Mining(DM)?

- Vaguely defined by several DM researchers depending on their background and views.

- Most common definition is "the process of uncovering previously unknown patterns and relationships in large databases using sophisticated statistical analysis and modeling techniques".

Multidisciplinary

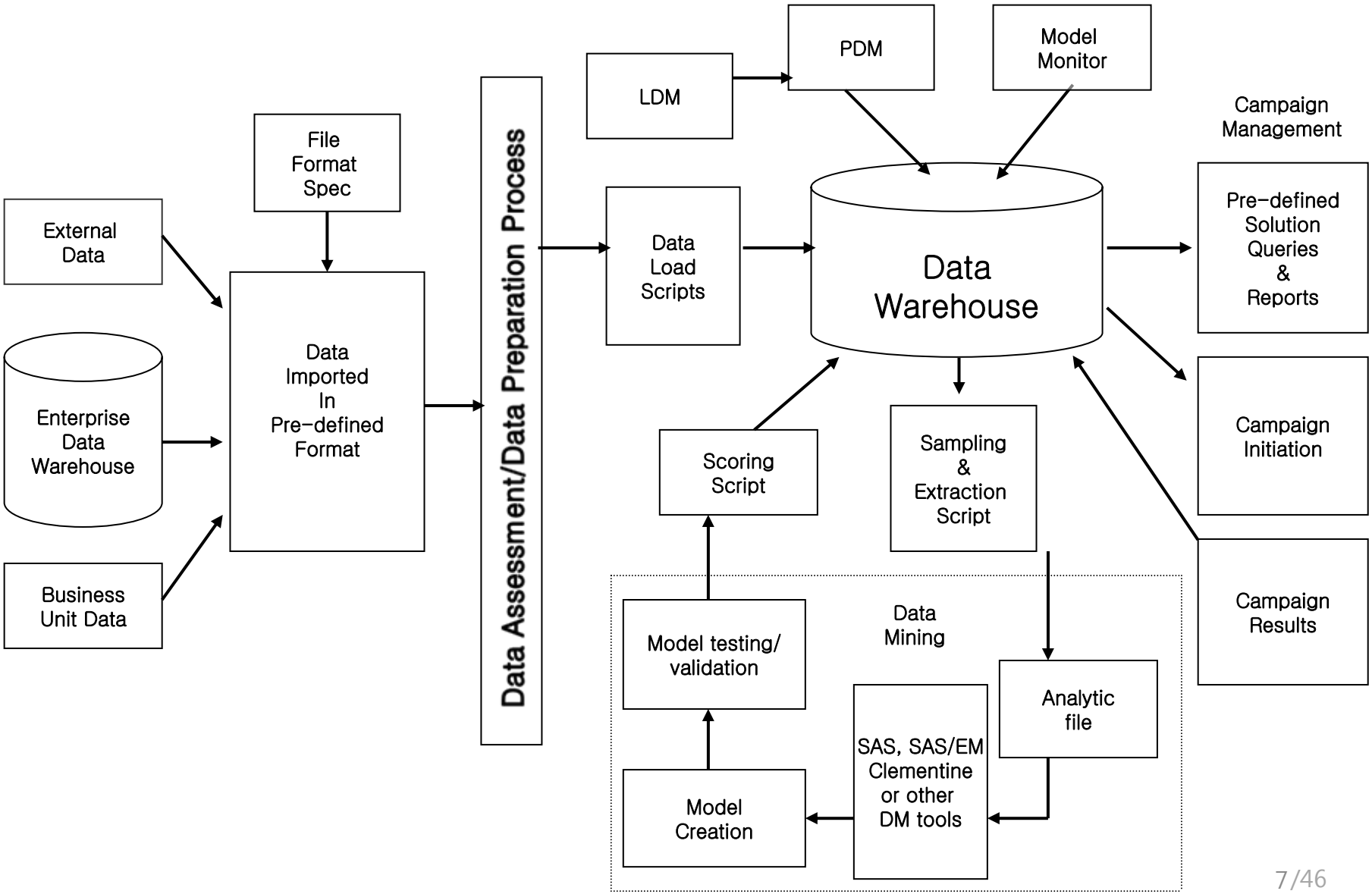


MIT 선정 10대 최신 기술

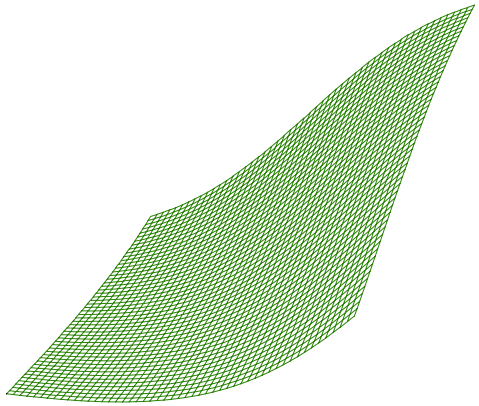
1. Brain-Matching Interfaces
2. Flexible Transistors
3. Data Mining
4. Digital Rights Management
5. Biometrics
6. Natural Language Processing
7. Microphotonics
8. Untanging Code
9. Robot Design
10. Microfluidics

From Technology Review, January/Februrary 2001

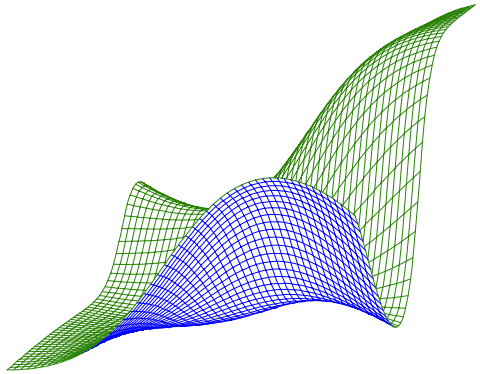
CRM Structure



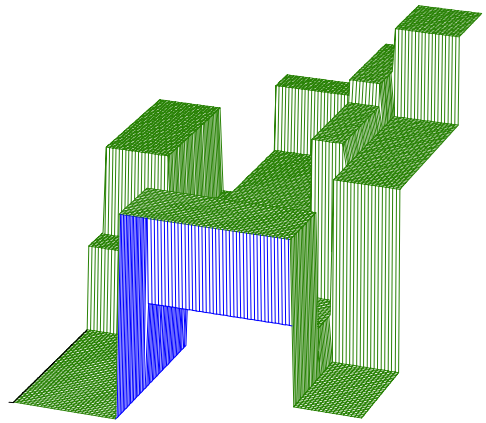
Modeling Methods



Generalized Linear Models



Neural Networks



Decision Trees

Data Mining Applications(General)

- Text Mining
- Pattern recognition
- Bio-informatics
- Web mining
- Financial engineering
- Marketing engineering
- Health studies
- Environmental engineering
- Spatial Data Mining

Data Mining Applications for Business problems.

- Customer Acquisition
- Cross-Selling/ Up-Selling
- Customer Retention (Reducing churn or attrition)
- Fraud Detection
- Customer Segmentation
- Market Basket Analysis(Association(Affinity) Analysis)
- Credit Risk Management(Credit Scoring)
- Customer Profitability Analysis
- Campaign Management
- Customer Life-Time Value Model (LTV)
- Customer Loyalty Analysis
- Healthcare Informatics (Medical Informatics)

Text Mining

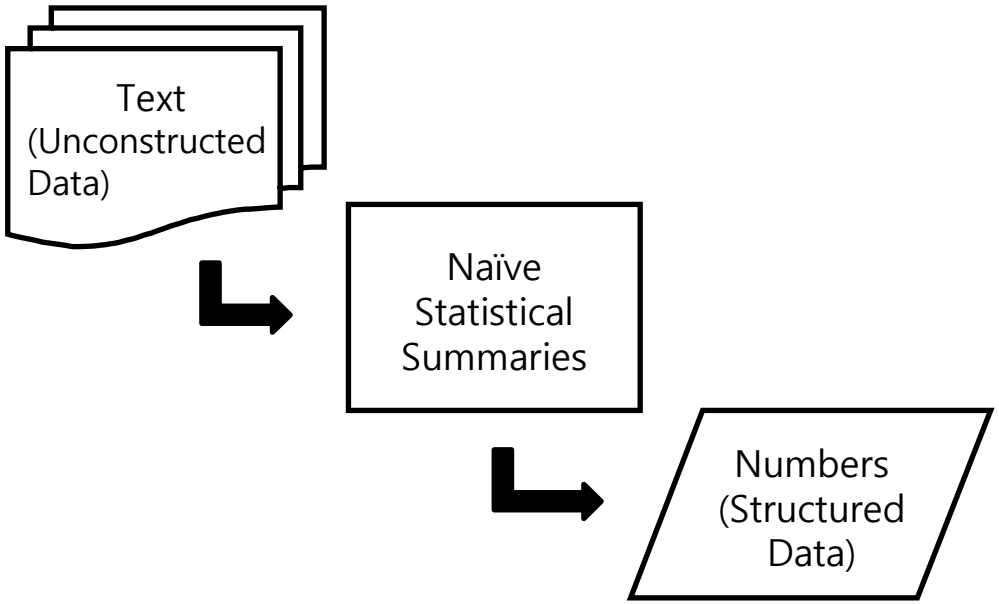
Text Mining

개념

- 비/반정형 데이터에 대하여 자연연어처리(Natural Language Process)기술과 문서 처리 기술을 적용하여 유용한 정보를 추출, 가공하는 목적으로 하는 기술
- A methods for extracting useful information from large and often unstructured collections of texts.

= information retrieval

Text Mining Goal: Convert Unconstructed data to Structured Data



Background

- 실생활에서 만들어지는 대부분의 자료는 문서 형태
  - 여러 분야의 논문
  - 신문 또는 잡지의 기사
  - 여론 조사, 콜센터의 전화 보고서
  - e-mail
- 인터넷의 발달
  - 디지털 형태의 문서
- 기존의 통계 분석이나 데이터 마이닝 기법을 적용하기에 부적합한 데이터 (문서형태의 unstructured data)

Now

Donga.com 2009.10.12 기사

이와 함께 SK텔레콤이 선보일 기술은 온라인 웹 페이지를 자동으로 한글로 번역해주는 '자동번역(머신 트랜슬레이션)'과, 기업 간 거래(B2B) 시장을 겨냥해 내놓는 '자연어 검색(텍스트 마이닝)'이 있다. 이 중 텍스트 마이닝은 기업이 소비자 선호도 조사 또는 마케팅 자료 조사 시 원하는 데이터만 검색해주는 시스템으로 지난달 '시맨틱 검색'을 선보인 SK커뮤니케이션즈와 공동으로 연구 중이다. 예를 들어 영화 '해운대'의 온라인 평가를 조사하기 위해 '해운대 영화 평가'로 검색하면 '해운대' '영화' '평가'란 말이 각각 들어 있는 모든 검색 결과가 아닌 사용자가 의도한 것만 제시하는 시스템이다. 이를 위해 SK텔레콤은 가격을 표현할 때, 맛을 평가할 때 쓰는 말 등 표현별 지식체계를 분류했으며 감정이나 선호에 대한 표현 역시 8만3000여 개의 DB를 만들었다.

SK텔레콤이 내년부터 선보일 서비스

	내용	주요 타깃
휴대전화 음성인식	말로 문자메시지 전송, 말로 무선인터넷 검색, 녹취 시 문서로 저장	휴대전화 가입자 시각장애인, 장년층
자동번역	해외 웹사이트를 수초 만에 한글로 번역	일반인
텍스트 마이닝 (Text Mining)	기업이 원하는 검색 결과만을 뽑아 주는 '맞춤형' 정보검색 서비스	기업

Text Mining

응용분야

- 과학논문 데이터베이스에서 특정 주제의 논문들을 찾아냄.
- 데이터마이닝 관점에서 문서로부터 구조화 된 정보를 추출하여 데이터 베이스화 시키거나 규칙을 찾아냄
- 사용자가 웹 상에서 문서를 찾는 것을 도와줌
- 사용자 프로파일의 생성 및 분석
- 대량의 DB에서 문서의 분류 및 군집화
- 문서 분류 정보를 이용한 문서 재해석
- 신문/논문/보고서 요약
- 문서 번역
- Spam filtering
- 문서 여과(filtering) 및 추천(recommendation)
- 대표적 키워드나 토픽의 추출
- 질의 응답 시스템
- 설문지 조사의 기타항목 요약
- Bioinformatics

Text Mining

용어

Terms: the contents of a document

Index: a list of all the terms in a document collection

Inverted index: for each term, a list of all document that contain that particular term.

Index를 만들기 전에 다음의 두 가지 perprocessing 단계가 필요함:

1) Stop words : words that one can find in any document. e.g. a, about, above,...
2) Stemming: the process of reducing each word that has a suffix to its stem.

   e.g.: **comput**able, **comput**ation, **comput**ing, **comput**ed, **comput**ational ➔ **comput**

Text Mining

- 통계적 분석 기법

  - 주어진 키워드(or query)에 근거하여 문서들을 할당하는 분류기법
    - 의사결정나무 (Decision Trees)
    - 신경망 분석(Artificial Neural Networks)
    - 베이지안 분류(Naïve Bayes Classifier)
    - 최근접 분류 (K-Nearest Neighborhood)
    - SVM(Support Vector Machine)
    - PCA (Principal Component Analysis)
    - Latent Semantic Indexing
    - 응용: 스팸 필터링

  - 사전 정보 없이 비슷한 문서들을 집단으로 묶는 군집기법
    - K-means
    - SOM
    - EM 군집

Data Mining vs. Text Mining *

비교 내용	데이터마이닝	텍스트마이닝
대상데이터	수치/범주화된 데이터	텍스트
데이터구조	관계형 DB	비정형/정형 텍스트
목표	미래 상황 결과의 예견/예측	적합한 정보를 획득하고, 의미를 정제하고 범주화함
방법	기계 학습	기계 학습 포함 인덱싱, 신경망 처리, 언어처리, 온톨로지 등 적용가능
성숙도	1994년 이후 광범위하게 구현	2000년 이후 광범위한 구현 시작

*From www.saltlux.com

Process

- Frequency weight
- Term weight

- SVD
- Roll up

- Hierarchical cluster
- EM cluster

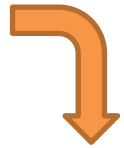
Text Mining _ process (1. Parsing)

- Parsing (문장의 분해)

Asymptotic / approximations / in / probability / and / statistics

Investigations / in / mathematical / statistics

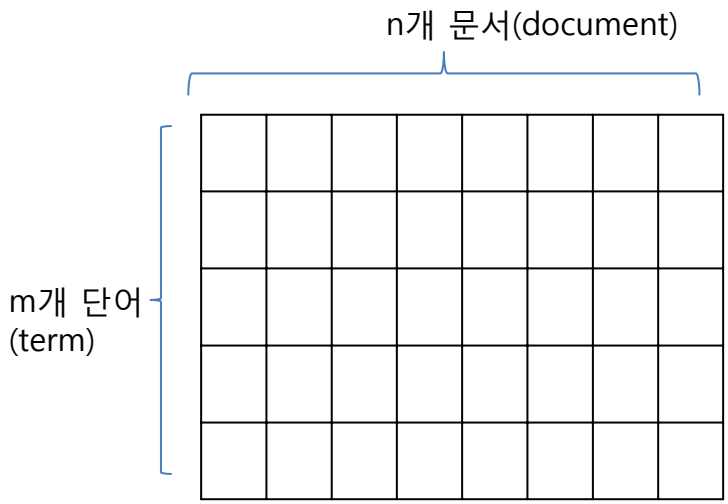
Probability / statistics / and / functional / analysis



	Doc1	Doc2	Doc3
Analysis	0	0	1
And	1	0	1
Approximations	1	0	0
Asymptotic	1	0	0
Functional	0	0	1
In	1	1	0
Investigations	0	1	0
Mathematical	0	1	0
Probability	1	0	1
Statistics	1	1	1

Unstructured data → Structured data

문서는 단어의 집합으로 이루어져 있다는 개념을 이용하여 행렬 형태로 나타냄 **(m*n term-document matrix)**

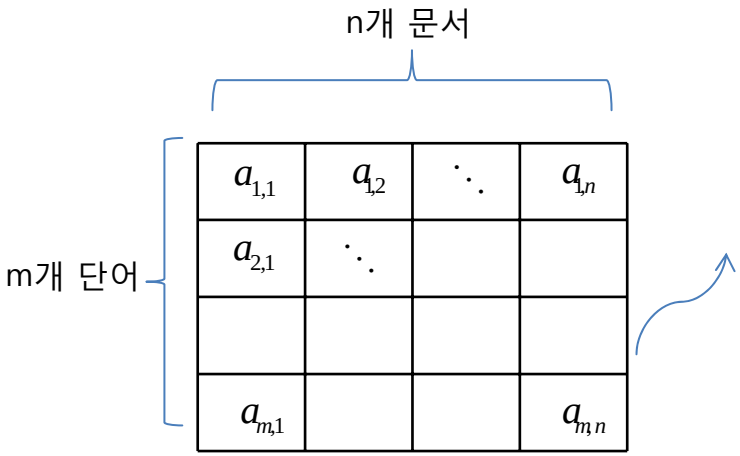


- 비구조화 데이터를 구조화 데이터로 바꿈
- 문서는 단어의 집합으로 이루어져 있다는 개념을 사용
- 각 셀 안에 들어가는 수는 문서에 포함된 단어의 빈도

Text Mining _ process (2. Weight)

Weighted Matrix

각 단어의 특징을 반영한 가중치를 사용하여 단어와 문서간에 상호 관계를 더 잘 나타내고자 함



$$a_{i,j} = L(i,j) \times G(i)$$

$L(i,j)$: 지역적 가중치 (Local Weight)

$G(i)$: 전역적 가중치 (Global Weight)

$a_{i,j}$ 는 문서(document) j 에서 단어 (term) i 의 가중된 도수(weighted frequency)를 말한다.

Text Mining _ process (2. Weight)

- 지역적 가중치 (Local Weight): L(i,j) 도수 가중치 (frequency weight)

- 빈도 $L(i, j) = tf(i, j) \quad tf: \text{term frequency}$
- 로그 $L(i, j) = \log(tf(i, j) + 1)$
- 이항 $L(i, j) = 1, \quad \text{if } tf(i, j) \geq 1$
  $L(i, j) = 0, \quad \text{if } tf(i, j) = 0$

지역적 가중치는 어떤 문서(document)가 전체 문서(corpus) 중에서 차지하는 정보를 반영하지 못한다. 즉, 문서가 클수록 상대적으로 작은 문서보다 더 큰 도수(frequency)를 가질 가능성이 커진다.

=> 이에 대한 해결책으로 단어 가중치(Term Weights = Global Weights)를 동시에 고려함.

Text Mining _ process (2. Weight)

- 전역적 가중치 (Global Weights= Term Weights)

  : 전체 문서 (Corpus)에서 어떤 단어(term)의 분포를 반영.

- Entropy

$$G(i) = 1 + \sum_j \frac{p_{ij} \log_2(p_{ij})}{\log_2(N)}$$

$G(i) = 0$ if $a_{i,j}$ is always 1,

$G(i) = 1$ if $a_{i,j}$ is 1 for only 1 document

- GF-IDF $G(i) = \frac{g_i}{d_i}$

- IDF $G(i) = \log_2\left(\frac{N}{d_i}\right) + 1$

- Normal $G(i) = \sqrt{\frac{1}{\sum_j L(i,j)^2}}$

- g_i : 전체 문서(corpus)에서 단어 i가 나타난 도수
- N : 전체 문서(corpus)에서 총 문서의 수
- d_i : 단어 i가 포함되어 있는 문서들의 수
- $a_{i,j}$: 문서 j에 단어 i가 나타난 도수

- $p_{i,j} = a_{i,j} / g_i$

GF-IDF : Global Frequency multiplied by Inverse Document Frequency

IDF: Inverse Document Frequency

Weighted term-document frequency matrix

	Document		
Term	D1	D2	Dn
T1	$\hat{a}_{1,1}$	$\hat{a}_{1,2}$	$\hat{a}_{1,n}$
T2	$\hat{a}_{2,1}$	$\hat{a}_{2,2}$	$\hat{a}_{2,n}$
Tm	$\hat{a}_{m,1}$	$\hat{a}_{m,2}$	$\hat{a}_{m,n}$

$\hat{a}_{ij} = G(i) * L(i,j)$ where $G(i)$ is term weight and $L(i,j)$ is frequency weight .

- Frequency weight와 term weight를 결합하면 문서 내, 문서 간 정보를 다 얻을 수 있다. 행렬 안에 들어가는 값들은 Frequency weight와 term weight의 곱으로 계산된다.

- 비구조화 텍스트 파일을 숫자의 배열로 이루어진 구조로 변환된 형태
- 거의 0으로 이루어져 있으며(sparse data), 실제로는 매우 크다라는 문제점
  - Matrix에서 row(낱말)의 차원을 줄이기 위해서 그 해결 방안으로 Transformation을 하게 된다.

Text Mining _ process (3. Transform)

Term Variables	Documents			
Term 1	Document1 1	Document2	⋯	Document n
Term 2	Data			
⋮				
Term m				

문제점 : (1) 차원이 크다
(2) Sparse → 대부분 0의 값을 가짐
단어의 수 (m) >> 문서의 수 (n)

=> 차원 축소 필요

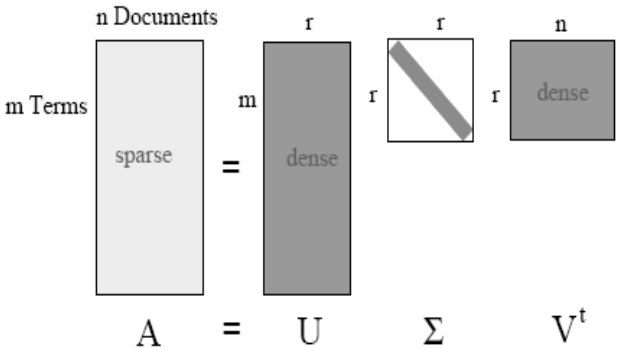
차원 축소(Dimension Reduction) 방법
1. 특이값 분해(SVD)를 통한 차원 축소
2. Roll up term을 이용한 차원 축소

Text Mining _ process (3. Transform)

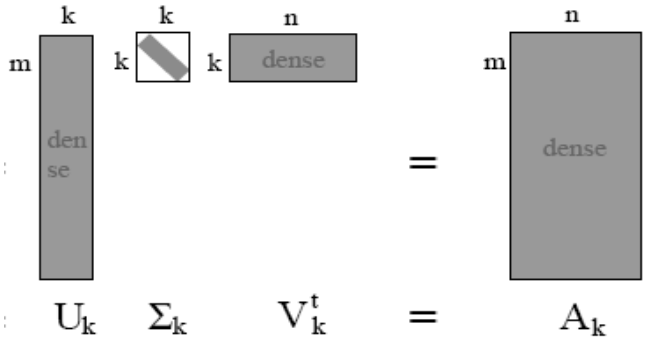
1. 특이값 분해(Singular Value Decomposition)를 이용한 차원 축소

$$A = U \Lambda V^T$$
$$m \times n \quad m \times m \; m \times n \; n \times n$$

Reduced SVD



Approximating



Text Mining _ process (3. Transform)

Dimension Reduction

문서들을 본래 정의된 단어 공간보다 훨씬 작은 차원에 나타내고자 함

Step1) D라는 단어-문서 행렬을 차원 축소하여 D_K라는 행렬을 만든다.

$$D \cong D_k = U_k \Sigma_k V_k^T$$

Step2) 양변에 U_K 의 전치행렬을 곱해준다.

$$U_k^T D_k = \Sigma_k V_k^T$$

step3) $U_k^T D_k$ 는 단어의 차원을 축소한 행렬로서 각 문서는 단어의 선형 결합으로 표현된다.

step4) step3)의 행렬에 데이터 마이닝 기법들을 적용하여 분석한다.

Text Mining _ process

SVD 갯수에 대한 Guideline

1. 분석목적이 데이터 탐색용(exploration)이라면 (e.g. Clustering) SVD 수는
   2개에서 50개 사이가 적당.

2. 분석목적이 데이터 예측용(prediction)이라면 (e.g. Classification) SVD 수는
   30개에서 200개 사이가 적당.

Text Mining _ process

Interpretation of the SVD

- An SVD projection is a linear combination of the values in a row or column of the term-document frequency matrix.
- A linear combination can be interpreted as an extension of the idea of a weighted average.
- **A weighted average of terms produces a concept**.
- Therefore, the SVD converts **terms** into **concepts**.

- SVD analysis is equivalent to an PCA of the data matrix, once the mean of each variable has been removed. Nonetheless, it is informative to look at dimensionality reduction from the SVD point of view, since it is not always desirable to remove the mean from data, especially if the data is relatively sparse.
- SVD 1 is the best linear combination of the term values.
- SVD 2 is the second linear combination of the term values.
- And so on...

실증 예제

1. Visualization
2. Query Matching & Classification (Latent Semantic Indexing)
3. Text Clustering(군집 분석)

Text Mining _ Example 1

- Document 1 -- deposit the cash and check in the bank
- Document 2 -- the river boat is on the bank
- Document 3 -- borrow based on credit
- Document 4 -- river boat floats up the river
- Document 5 -- boat is by the dock near the bank
- Document 6 -- with credit, I can borrow cash from the bank
- Document 7 -- boat floats by dock near the river bank
- Document 8 -- check the parade route to see the floats
- Document 9 -- along the parade route.

	d1	d2	d3	d4	d5	d6	d7	d8	d9
the	2	2	0	1	2	1	1	2	1
cash	1	0	0	0	0	1	0	0	0
check	1	0	0	0	0	0	0	1	0
bank	1	1	0	0	1	1	1	0	0
river	0	1	0	2	0	0	1	0	0
boat	0	1	0	1	1	0	1	0	0
+ be	0	1	0	0	1	0	0	0	0
on	0	1	1	0	0	0	0	0	0
borrow	0	0	1	0	0	1	0	0	0
credit	0	0	1	0	0	1	0	0	0
+ float	0	0	0	1	0	0	1	1	0
by	0	0	0	0	1	0	1	0	0
dock	0	0	0	0	1	0	1	0	0
near	0	0	0	0	1	0	1	0	0
parade	0	0	0	0	0	0	0	1	1
route	0	0	0	0	0	0	0	1	1
parade route	0	0	0	0	0	0	0	1	1

Doc 1,3, and 6 : about banking (은행) at a financial institution

Doc 2,4,5,and 7 :about the bank(제방) of a river

Doc 8 and 9: about the parade

* SAS Text miner help documentation

"Check" as noun in Doc 1, verb in Doc 8 (different meaning)

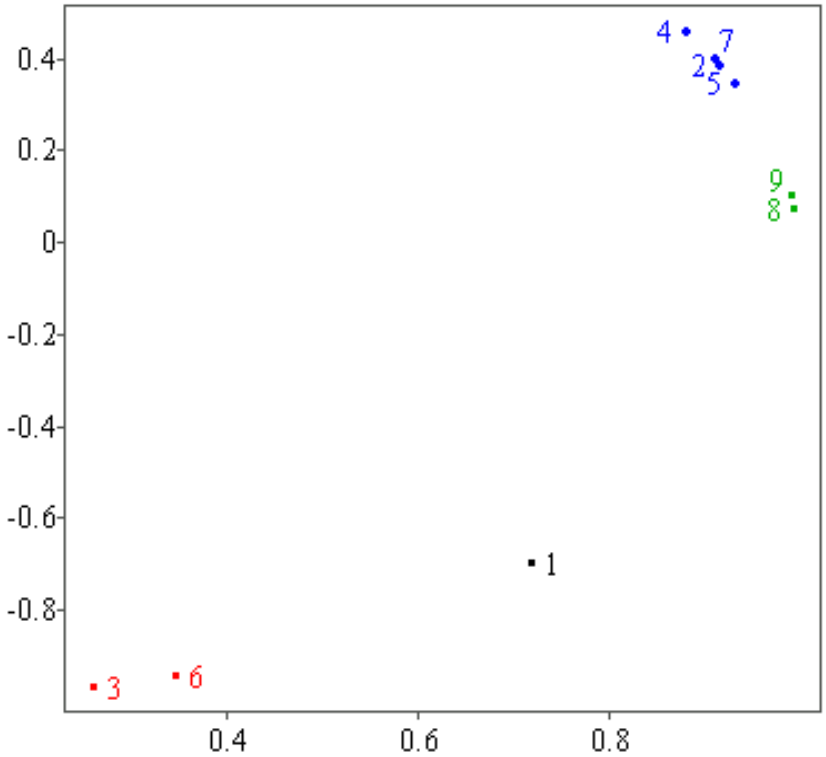
"Floats" as verb(떠 있다) in Doc 4, an object (장식차량(꽃수레)) in Doc 8.

Doc 1 and 2 are not related even if they share the same word "bank", but doc1 and 3 are similar.

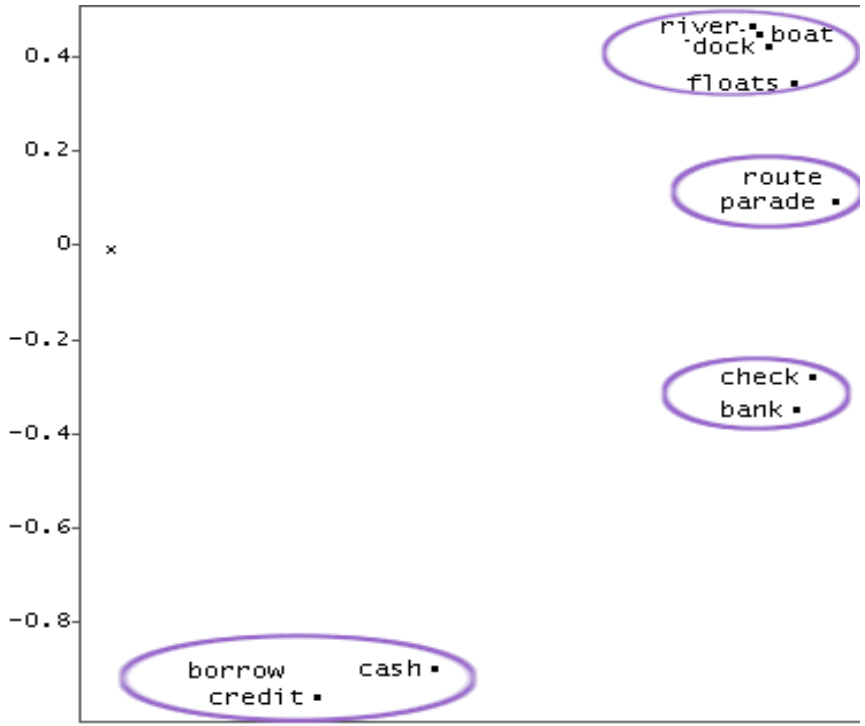
=> Use SVD

Text Mining _ Example 1

Scatter plot of document by SVD



Scatter plot of term by SVD



Doc 1 is closer to Doc 3 than Doc 2 even though Doc1 and 3 do not share any of the same words.

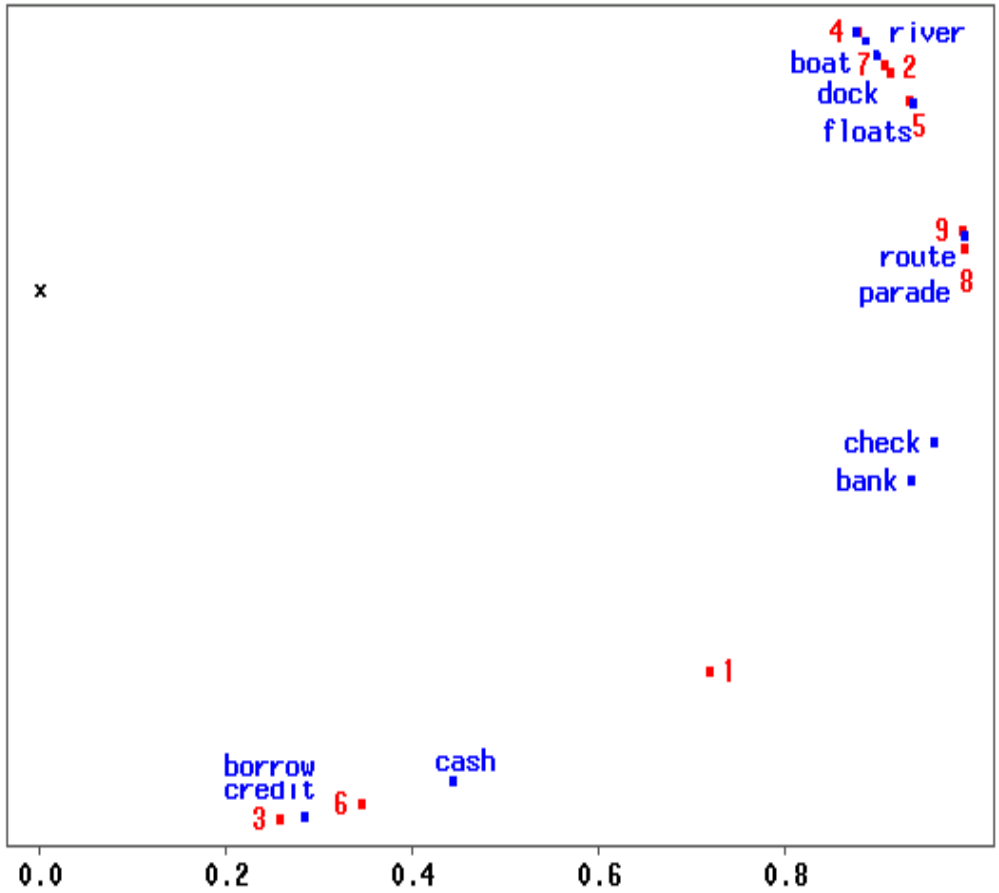
Doc 5 is related to Doc 2,4, and 7.

* SAS Text miner help documentation

SVD represents term with 2 dim rather than the original 11 dim.

Text Mining _ Example 1

Scatter plot of documents and terms all together



* SAS Text miner help documentation

Text Mining _ Example 2

Query Matching & Classification (Latent Semantic Indexing)

SVD 기법의 Text Mining에의 활용

Document 1 : The **Google™ matrix** P is a model of the **Internet**.
Document 2 : is nonzero if there is a **link** from **Web page** j to i.
Document 3 : The **Google matrix** is used to **rank** all **Web pages**.
Document 4 : The **ranking** is done by solving a **matrix eigenvalue** problem.
Document 5 : **England** dropped out of the top 10 in the **FIFA ranking**.

*Key word 는 굵은 글씨

Query (q) : "**ranking** of **Web pages**"

Text Mining _ Example 2

Term	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
eigenvalue	0	0	0	1	0
England	0	0	0	0	1
FIFA	0	0	0	0	1
Google	1	0	1	0	0
Internet	1	0	0	0	0
link	0	1	0	0	0
matrix	1	0	1	2	0
page	0	1	1	0	0
rank	0	0	1	1	1
Web	0	1	1	0	0

Query (q) : "**ranking** of **Web pages**"

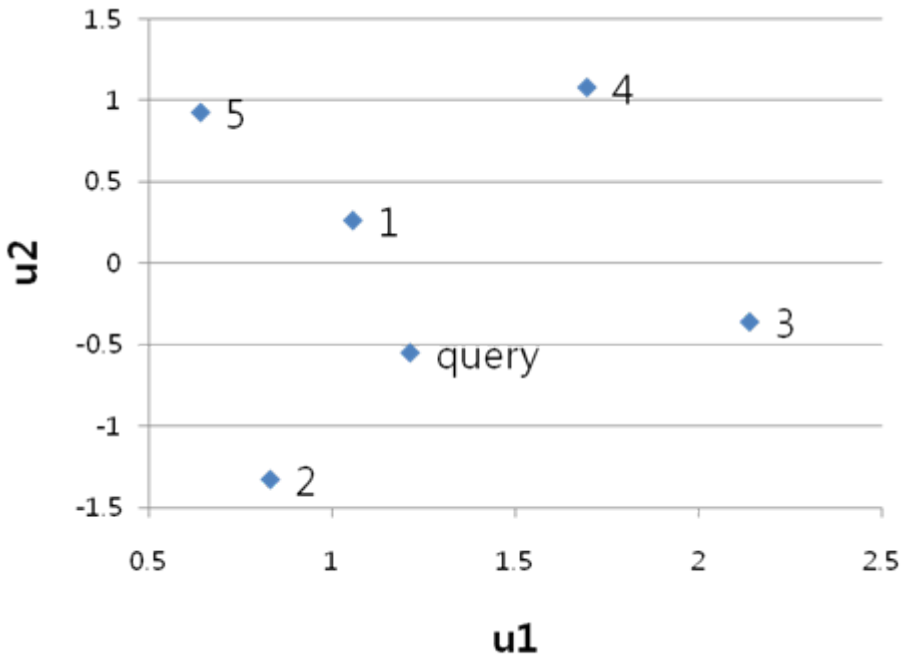
Cosines for the query and the original data
(0 0.6667 0.7746 0.3333 0.3333)

After Compute SVD
(0.7857 0.8332 0.9670 0.4873 0.1819)

Note: cosine distance measure: $\cos(\theta(q, a_j)) = \dfrac{q^T a_j}{\| q \|_2 \| a_j \|_2}$

Text Mining _ Example 2

u1	u2
0.1425	0.243
0.0787	0.2607
0.0787	0.2607
0.3924	-0.0274
0.1297	0.074
0.102	-0.3735
0.5348	0.2156
0.3647	-0.4749
0.4838	0.4023
0.3647	-0.4749



u1 :Strong values in Google, matrix, etc..

Text Mining _ Example 3 _ 군집 분석

Data

1990년부터 2003년까지의 미국 국가 과학 재단(National Science Foundation, NSF)으로부터 상을 받은 연구 요약(abstract) 자료를 사용

41,717개의 연구 제목을 가지고 비슷한 연구끼리 군집해 보고자 함.

Tool

SAS Text Miner version 3.2

Weight combination

Log-entropy

Cluster method

EM cluster

Text Mining _ Example 3 _ 군집 분석

Determine number of cluster

1. Average RMS std .(Root mean square standard deviation)

$$\sqrt{\frac{w_K}{d(N_K - 1)}}$$

2. Pseudo F statistics

$$F^*(K) = \frac{\frac{BGSS(K)}{K-1}}{\frac{WGSS(K)}{N-K}}$$

3. Hartigan index

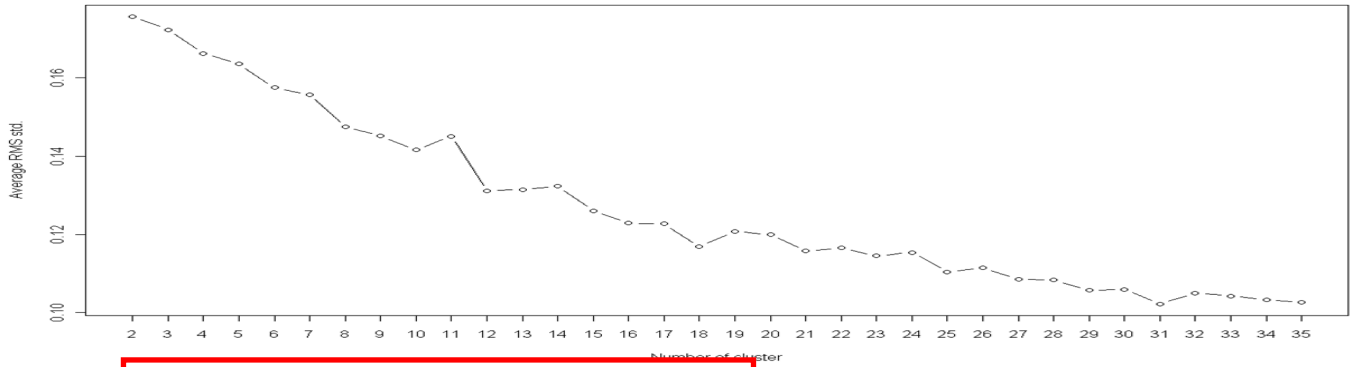
$$H(K) = \left[\frac{WGSS(K)}{WGSS(K+1)} - 1\right](N - K - 1)$$

w_K	군집 개수
d	K 군집에 있는 각 문서들의 군집의 평균으로부터 거리 제곱합
N_K	차원의 수
N	군집 K에 있는 문서의 수
	총 문서의 수
BGSS(K)	각 문서 군집의 평균과 전체 평균 간의 거리 제곱합
WGSS(K)	각 문서와 그 문서가 속한 군집 평균과의 거리 제곱합

Text Mining _ Example 3 _ 군집 분석

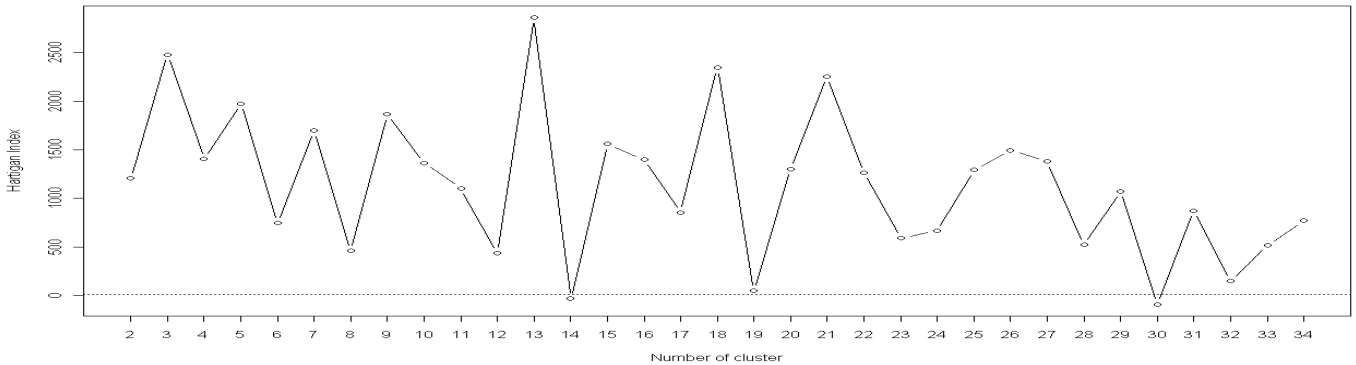
Determine number of cluster

Average
RMS std.



Pseudo
F statistic

Hartigan
Index



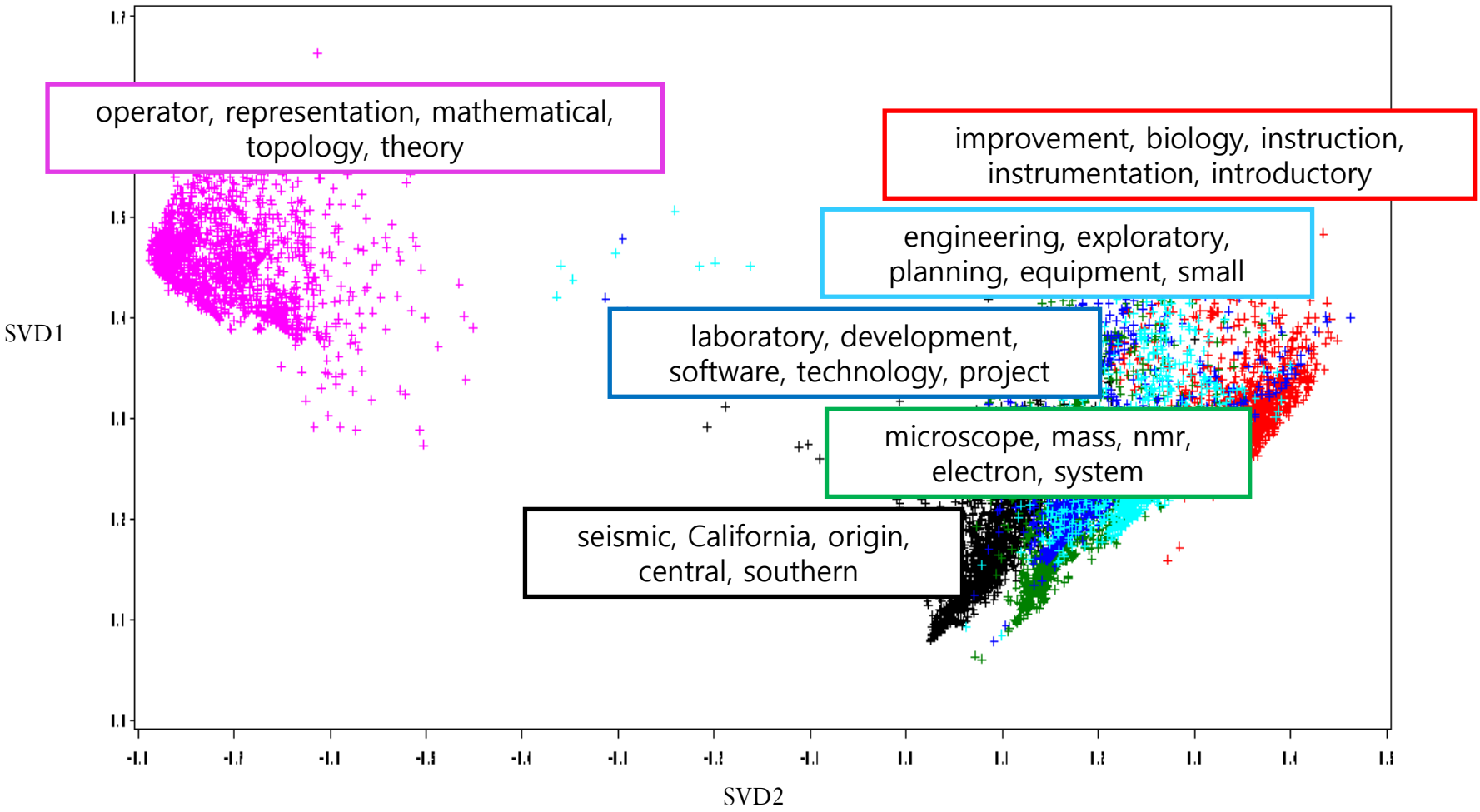
Text Mining _ Example 3 _ 군집 분석

SVD

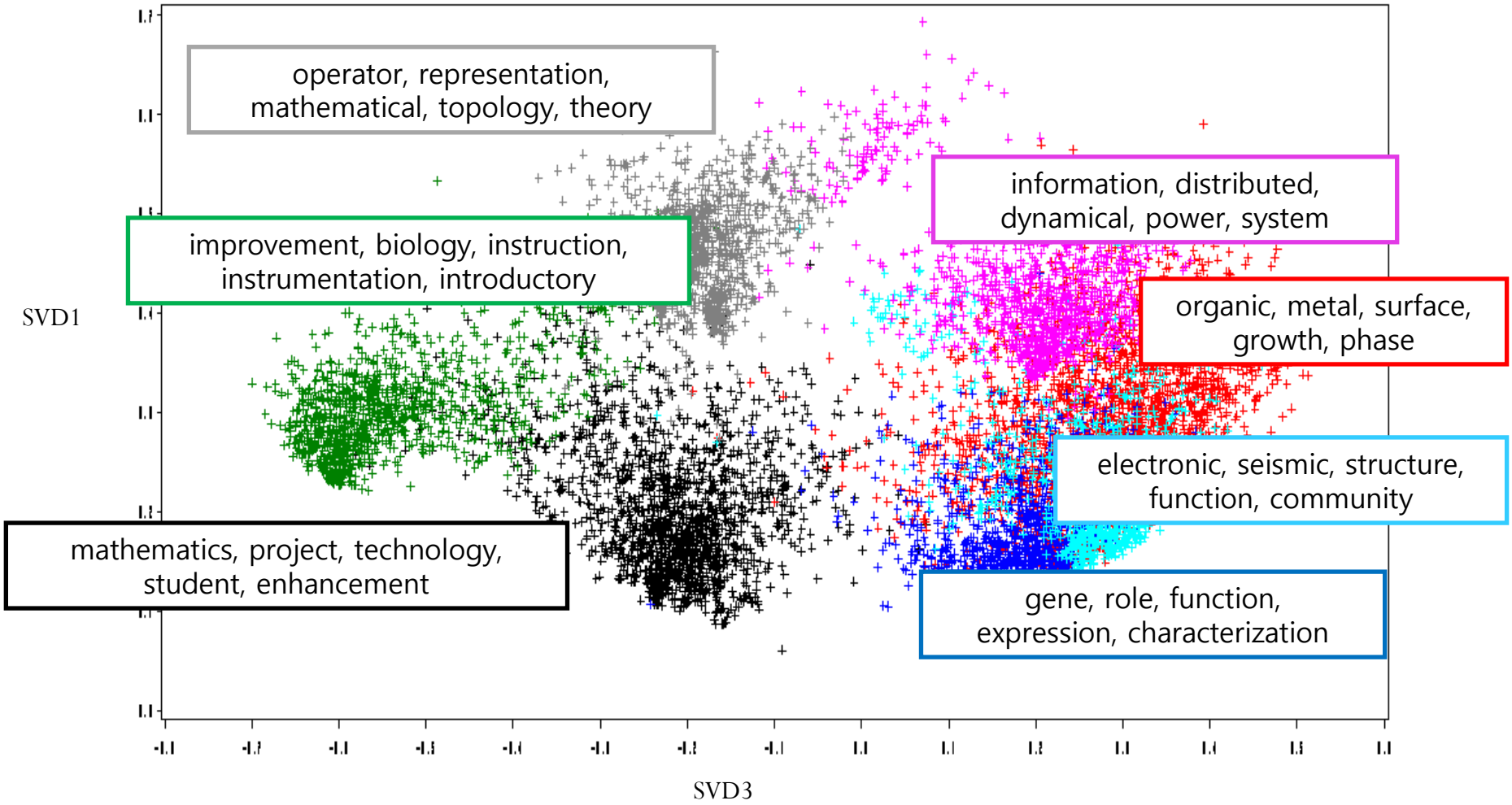
단어의 선형 결합으로 이루어지는 SVD에서 각 단어의 계수의 상위 10개와 하위 10개를 뽑아 보았다.
SVD는 상위 3개만 나타내었다.

(-) SVD1	(+)	(-) SVD2	(+)	(-) SVD3	(+)
sister-chromatid	real	sum	electronics	revised	evaporation
plastids	other	diophantine	enhanced	major	agglomerate
segregation	reduction	mathematical	video	introductory	mobility
pneumoniae	dimensional	yang-mills	computer-based	miniority	transient
males	complex	noetherian	improved	computer-based	multicomponent
candida	multiple	differential equations	major	micro computer	lateral
trait	expansion	equations	support	laboratory	distribution
decline	domain	invariant	instructional	unix-based	copolymer
factorial	vector	asymtotic	theme	ample	compacted
gmp	stability	non-commutative	capability	inquiry	long

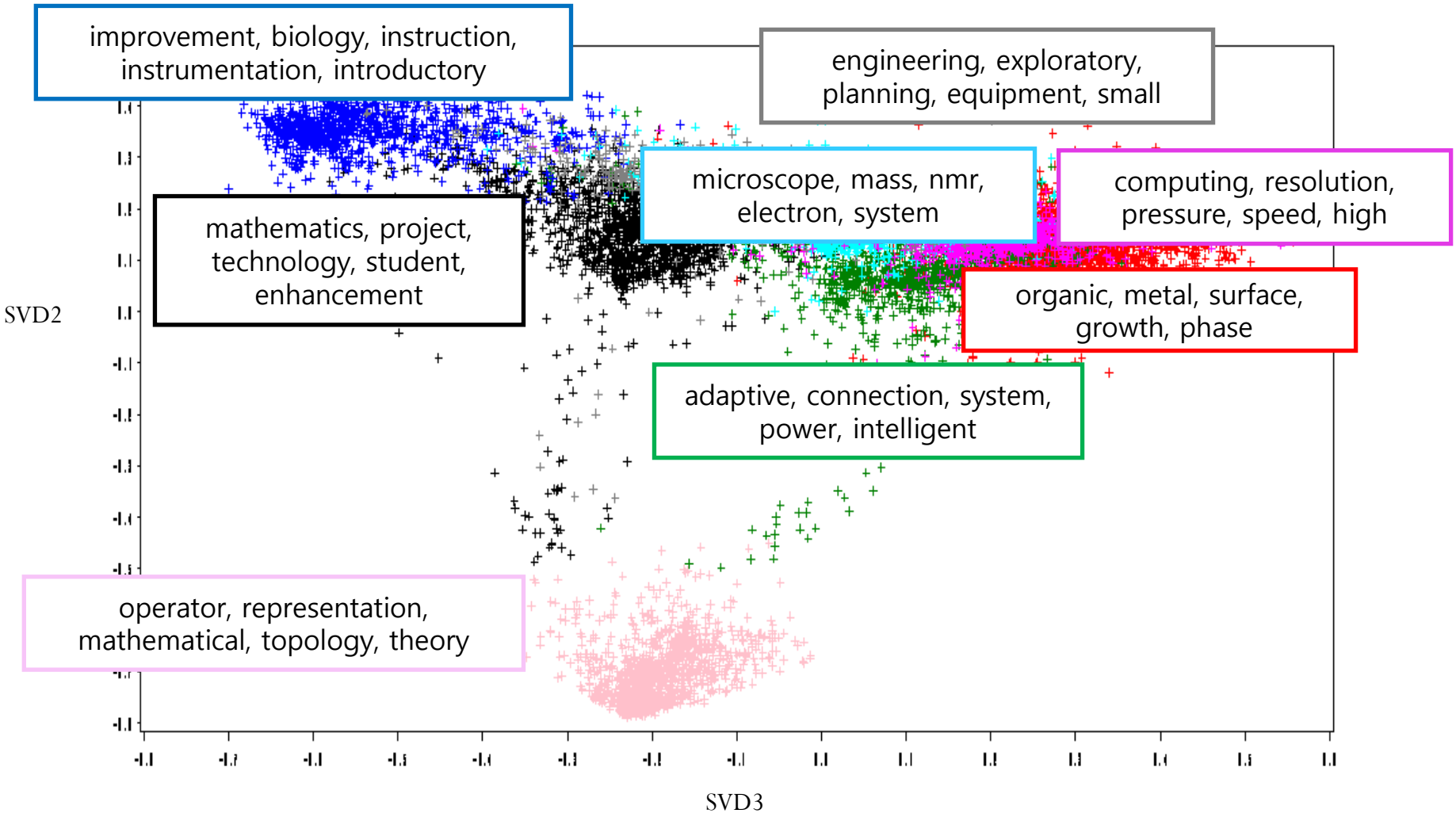
Text Mining _ Example 3 _ 군집 분석



Text Mining _ Example 3 _ 군집 분석



Text Mining _ Example 3 _ 군집 분석



Text Mining _ Example 3 _ 군집 분석

결론

각 군집은 5개의 단어를 통해 군집의 특성을 파악할 수 있다. (사람의 주관적인 해석이 포함됨.)

SVD1: 수학적인 성격을 가지는 군집 구분
SVD2: 수학 이론과 그 외의 군집을 구분하는 축
SVD3: 이과계통과 공학 분야의 논문 구분

각 축에 따라서 군집의 특징이 나누어짐을 확인할 수 있다.

나머지 SVD 축에 대해서 그림을 그리면 또 다른 형태의 그래프를 얻을 수 있을 것이며 그를 통해서 어떤 분야의 논문들이 가깝게 위치하는지 파악할 수 있을 것이다.

Text Mining _장,단점

장점/단점

- 자연어라는 것은 수 많은 종류로 표현이 가능하므로 분석 이전에 전처리 작업을 잘 해 주어야지 좋은 결과를 얻을 수 있음.
- 텍스트 분석의 전문가 이자 인도 방갈로르에서 K프락시스(K-Praxis)라는 웹사이트를 운용하고 있는 마단 판딧
  - "텍스트마이닝이 사람들로 하여금 산더미 같은 문서를 효과적으로 검색, 살아있는 정보검색을 가능하게 해줄 것"
  - "하지만 텍스트마이닝은 언어의 뉘앙스까지는 잡아 내지 못하는 단점을 안고 있다"

감사합니다.

Questions and Answers

yung@dongguk.edu