미래 유망 기술 발굴을 위한 텍스트 마이닝 기법

전종준 | 서울시립대 통계학과
문상준 | 서울시립대 통계학과
유형곤 | 안보경영연구원
신동협 | 안보경영연구원
김단비 | 안보경영연구원

2017. 12. 27

CONTENTS

텍스트 마이닝 소개

〈텍스트 마이닝〉

비정형 텍스트 데이터에서
자연 언어 처리 기술에 기반하여
유용한 정보를 추출, 가공하는 기법

출처 : Wikipedia

문서 및 텍스트 데이터를 이용하여
특정한 정보를 추출하는 작업

〈예시〉

- 키워드 빈도 및 추세 분석
- 주제어 분석
- 문서 및 연관검색어 추천
- 문서 요약
- 문장 생성

〈유용성〉

문장, 문서에 대한
정성적 분석의 대안

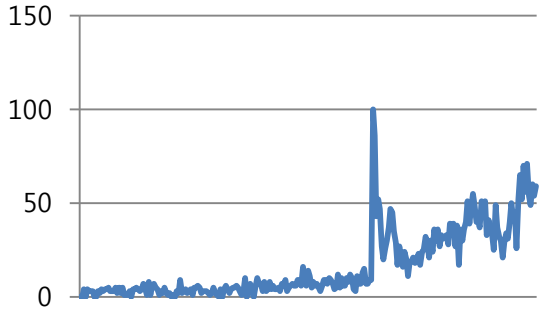
기술적 자원의 재활용성

기술 변화 예측에의 텍스트 마이닝의 활용

〈기술 변화 예측〉

과학 기술의 발전방향에 대한
중요성이 증가하여

기술 변화에 대한 예측을 통해
미래변화 전망 및 수요를 도출



인공지능 키워드 출현 추이

출처 : 구글 트렌드

〈텍스트 마이닝 활용〉

- 기술변화 예측에 필요한 정보 증가
- 특허 문서의 네트워크 분석 방법론 개발
- 특정 기술 분야의
  키워드 기반 기술 변화 예측
- 대용량 데이터를 이용한
  텍스트 마이닝 기법 적용

〈관련 문헌 연구〉

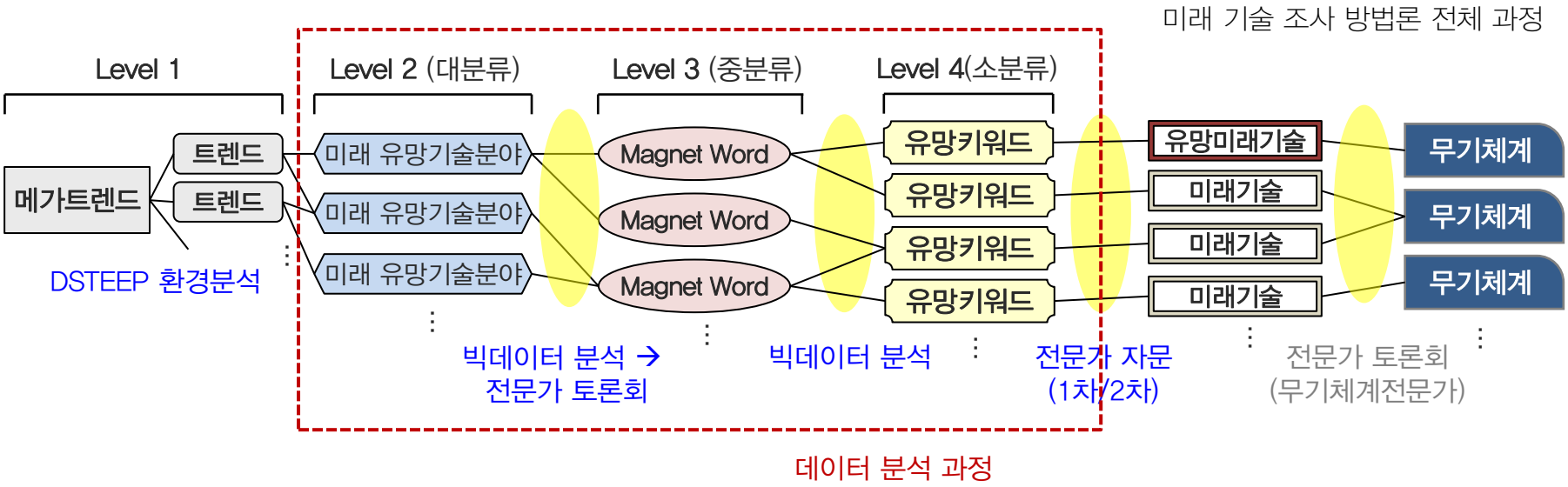
- Zhu and Porter (2002) :
  전문 자료원과 일반 자료원에서 출현한
  키워드 빈도수를 비교하여
  기술 혁신성 정의 및 시각화
- Yoon and Park (2004):
  특허 문서의 키워드 벡터와 상관행렬을 이용한
  문서 중심도 계산
- Chang et al. (2009):
  특허 데이터의 인용관계를 이용하여
  거리를 계산하고 클러스터링 방법을 적용
- Ho et al. (2014):
  연료전지를 주제로 한 논문의 트렌드 분석과
  논문 주제의 정성적 분석 수행

분석 목표

과학 기술에 대한 텍스트 데이터와
미래 기술 변화의 예측에 대한 수요가 증가했다.

따라서 과학 기술에 대한 문헌 정보와 텍스트 마이닝 기법을 이용하여
미래 기술 변화에 대해 예측하고 과학기술 트렌드의 변화를 분석한다.

전체 미래 기술 조사 과정



미래 유망기술분야(대분류)로부터 Magnet word를,
Magnet word로부터 유망기술키워드를 도출할 때
데이터 분석 기법을 활용

각 데이터 분석 과정

데이터 수집

텍스트 크롤링 기법을 이용
학술지 사이트의 초록 데이터 수집

데이터 전처리

자료원 제한
단어 사전 데이터와 결합
형태소 분석
키워드 정제

추출

분석에 필요한
데이터 생성

데이터 분석

연관성 분석

Graphical lasso

분석 결과

데이터 분석 결과 도출
시각화

데이터 수집

〈학술지 데이터 수집〉

- 웹 문서의 규칙에 따라 데이터를 수집하는 **"텍스트 크롤링"** 기법 사용
- 학술지 사이트별로 개별적인 데이터 수집 프로그램 작성
- 데이터 수집 프로그램은 크게 API 제공 유무에 따라 분류

- 입력 : 키워드, 연도, 자료원
- 출력 : 제목, 저자, 초록 등 관련 문헌 정보

〈데이터 수집 결과 예시〉

source	url	title	date	issn	abstract
1	http://ieee	Magnetic	2016	0278-0062	magnetic particle imaging mpi is able to provide high temporal and good spa
1	http://ieee	Semiautor	2015	0196-2892	this paper proposes a novel algorithm for extracting street light poles from ve
1	http://ieee	Switchable	2015	2156-342X	a terahertz amplitude switching device is proposed which allows for the effici
1	http://ieee	Nonortho	2017	0018-9480	design of the power ground layout of a multilayered printed circuit board pcb
1	http://ieee	CMOS Mi	2017	0018-9200	a flow cytometer chip fabricated in nm standard cmos technology embedded
1	http://ieee	Effect of E	2015	2156-3381	dielectric nanoparticle arrays have been proposed as antireflection coatings a
1	http://ieee	Crosstalk	2016	0018-9375	in this paper the concepts of weak coupling and weak imbalance are exploite
1	http://ieee	Radioluce	2016	0278-0062	four dimensional d ultrasound us is an attractive modality for image guidanc
1	http://ieee	Transient	2015	0018-9375	an equivalent circuit model for the transient analysis of through silicon vias t
1	http://ieee	An INSPEC	2015	0018-9456	noncontact optical imaging is frequently used in the inspection and metrolog
1	http://ieee	Printed M	2017	0018-9480	although wireless sensor networks wsns have been an active field of research
1	http://ieee	3-D Inkjet	2017	1536-1225	the gain of an antenna can be enhanced through the integration of a lens al
1	http://ieee	Smart Opt	2017	1530-437X	proposed is a smart single viewing axis optical laser line illumination based c
1	http://ieee	Force Ripp	2015	0018-9464	a novel direct drive degree of freedom dof planar levitating synchronous mot
1	http://ieee	Integrated	2015	0093-9994	this paper explores the use of gan power fets to realize an integrated modula
1	http://ieee	Successive	2017	0018-926X	we propose a novel heuristic method for optimizing planar pixel antennas wh
1	http://ieee	V-Band Vi	2015	1531-1309	in this letter a grounded coplanar waveguide to microstrip gcpw to ms transit
1	http://ieee	Robust Sc	2015	1556-6013	vulnerability of iris recognition systems remains a challenge due to diverse pr
1	http://ieee	A Fluidic (	2015	1530-437X	a new sensor design for detecting water hardness using complexometric and

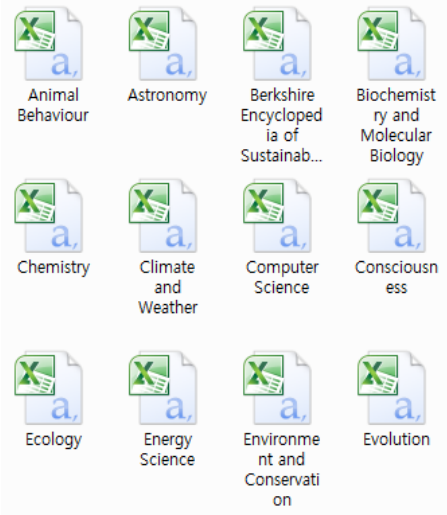
3D printing 분야 데이터 수집 결과

데이터 수집

〈단어 사전 데이터 수집〉

- 단어 정제를 위해 아래 사전의 단어를 수집하여 저장
  - IEEE taxonomy
  - Oxford dictionary (약 180,000 단어)
  - Taxonomy for Service Computing
  - NASA Technology Roadmaps TA 1: Launch Propulsion Systems
  - EDA Technology Taxonomy
  - NASA SBIR-STTR Technology Taxonomy
  - Space Technology Roadmaps

〈사전 데이터 수집 결과 예시〉



Oxford 사전 데이터 예시

데이터 전처리

〈자료원 제한〉

- 특성 세부 분야의 키워드 도출을 위해 자료원 제한
- 세부 학술지 별 문서 빈도를 계산
- 전문가의 의견을 반영하여 자료원 선정

3D printing 분야 제한된 자료원 예시

IEEE SENSORS JOURNAL
ONCOGENE
IEEE Transactions on Components Packaging and Manufacturing Technology
IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION
NATURE MATERIALS
IEEE Access
SCIENCE
IEEE TRANSACTIONS ON MEDICAL IMAGING
NATURE
JOURNAL OF MICROELECTROMECHANICAL SYSTEMS
IEEE Antennas and Wireless Propagation Letters
IEEE TRANSACTIONS ON ELECTROMAGNETIC COMPATIBILITY
NEUROPSYCHOPHARMACOLOGY
Nature Nanotechnology
Journal of Display Technology
3D printing and additive manufacturing

데이터 전처리

〈형태소 분석〉

- 문서를 단어 단위로 분해하여 정형 데이터로 처리
- 하나의 단어와 주변 단어들을 묶어서 복합어 생성
- 분석 시네 주변 6개 단어를 하나로 묶는 "6-gram" 사용

N = 1 : 이 문장은 예시 문장입니다.

N = 2 : 이 문장은 예시 문장입니다.

N = 3 : 이 문장은 예시 문장입니다.

예시 문장에 대한 3-gram 모형

〈키워드 정제〉

- 사전에 포함된 단어를 추출하여 문서-단어 행렬 생성
- 분야의 특성이 반영되지 않은 일반어의 필터링 필요
- 단어 출현의 유사성 측도인 Hellinger distance를 사용하여 단어 제거

keyword	value
gain	0.005155
modeling	0.011498
logic	0.012446
engineering	0.013475
simulation	0.013663
testing	0.014342
hardware	0.015504
guidelines	0.015717
stem	0.016377
art	0.018546
Analysis	0.024259
monitoring	0.025558

Hellinger distance 예시

데이터 분석 – 연관성 분석

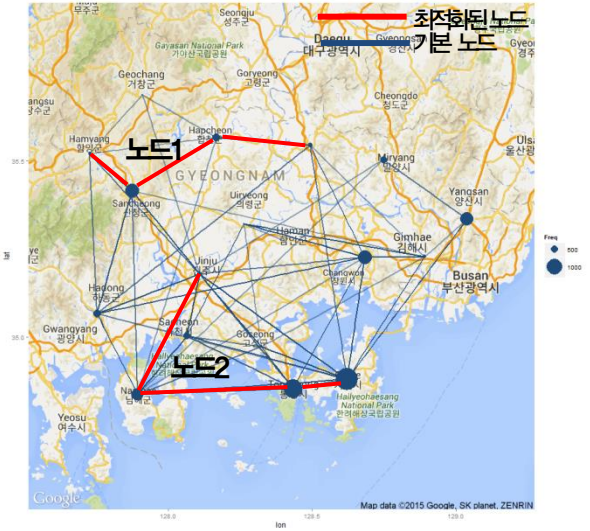
〈분석 방법〉

- 대규모 데이터베이스에서 변수 간의 상관성 발견을 위해 사용
- 하나의 단어에 대해 다른 단어가 출현하는 사건을 수치화
  - 지지도 (Support) :
    단어가 전체 문서에 포함된 비율
  - 신뢰도 (Confidence) :
    입력 단어로 검색 시 단어가 동시에 출현할 확률
  - 향상도 (Lift) :
    단어가 동시 출현할 빈도와 일반적인 출현 빈도의 비율

	지지도	신뢰도	향상도
{경남 통영시,문화} => {경남 거제시}	0.348432	0.775194	1.332219
{경남 사천시,문화} => {경남 남해군}	0.114983	0.673469	2.611969
{경남 고성군,문화} => {경남 거제시}	0.097561	0.7	1.202994
{경남 통합창원시,문화} => {경남 김해시}	0.069686	0.625	5.275735
{경남 거제시,경남 통합창원시,문화} => {경남 김해시}	0.034843	0.833333	7.034314

〈분석 목적〉

- 분석에서 신뢰도가 증가하며 향상도가 높은 단어를 선택
- 신뢰도를 통해 단어의 동시 출현 빈도 및 증가 추세의 단어를 선택
- 향상도를 통해 일반적인 출현 빈도보다 동시 출현 빈도가 높은 단어를 선택하여 일반적인 단어 제외



연관성 분석 예시

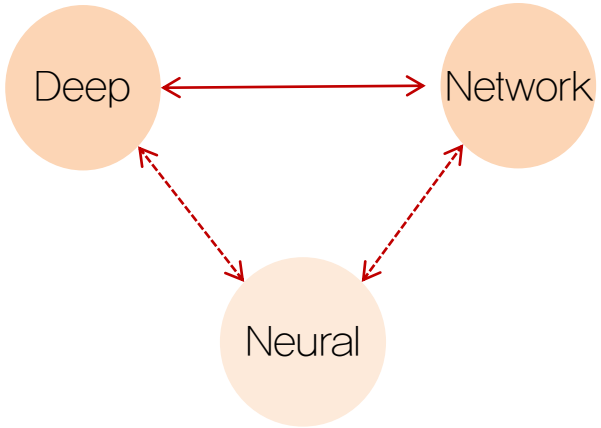
데이터 분석 – Graphical lasso

〈분석 방법〉

- Graphical Lasso 모형은 Gaussian graphical 모형의 확장
- Gaussian graphical 모형
  - 하나의 단어가 주어졌을 때 각 단어와 다른 단어의 상관성 편상관계수 추정
  - 편상관계수는 회귀분석을 통해 계산된 회귀계수로 파악
  - 회귀계수가 0인 경우 편상관성이 없다고 판단
- Graphical lasso 모형
  - 회귀계수를 0으로 만들기 위해 Lasso 벌점 함수를 부여

〈분석 목적〉

- 하나의 입력 단어에 대한 유의미한 단어들을 찾음
- 유의미한 단어들 사이의 상관성을 동시에 파악



Graphical lasso 모형 예시

분석 결과 – Magnet word 도출 결과

〈분석 방법〉

- 대분류로부터 Magnet word를 도출하기 위해 **연관성 분석 기법**을 사용
- 도출된 결과와 전문가 토론을 바탕으로 Magnet Word를 도출

〈분석 결과〉

단어	연도	신뢰도	향상도
machine learning	2017	0.158093	26.92673
neural networks	2017	0.099122	25.76233
artificial intelligence	2017	0.092848	48.85746
decision making	2017	0.031368	11.45673
supervised learning	2017	0.026349	17.16796
computer vision	2017	0.025094	14.12084
feature extraction	2017	0.025094	10.33714
internet of things	2017	0.022585	76.98925
machine learning algorithms	2017	0.017566	30.6056
pattern recognition	2017	0.017566	8.716786
artificial neural networks	2017	0.016311	24.59379
image analysis	2017	0.015056	10.50892
neuroscience	2017	0.012547	7.433891
linear discriminant analysis	2017	0.011292	26.82959
classification algorithms	2017	0.011292	25.5397
support vector machines	2017	0.011292	12.18408
complex networks	2017	0.011292	11.91089
gaussian mixture model	2017	0.010038	590.2509

인공지능 분야 결과 예시

분석 결과 – 유망 키워드 도출 결과

〈분석 방법〉

- Magnet word로부터 유망키워드를 도출하기 위해 **연관성 분석 기법, Graphical lasso**을 사용
- 도출된 결과와 전문가 토론을 바탕으로 최종 연관 키워드를 도출

〈분석 결과〉

context awareness

internet of things, **education**, **decision making**, heterogeneous networks, **base stations**, business, **centralized control**, data collection, encryption, **neural networks**

Graphical lasso 분석 결과

context awareness

machine learning algorithms, decision trees, informatics, data mining, artificial intelligence, **optimization methods**, mental disorders, centralized control, data privacy, classification algorithms, **base stations**, encryption, **training**

연관성 분석 결과

context awareness

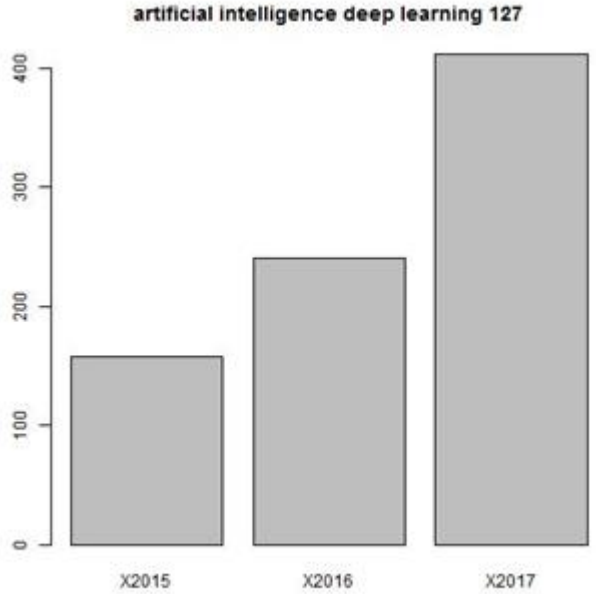
centralized control, **base stations**, encryption

두가지 분석방법 동시 도출 결과

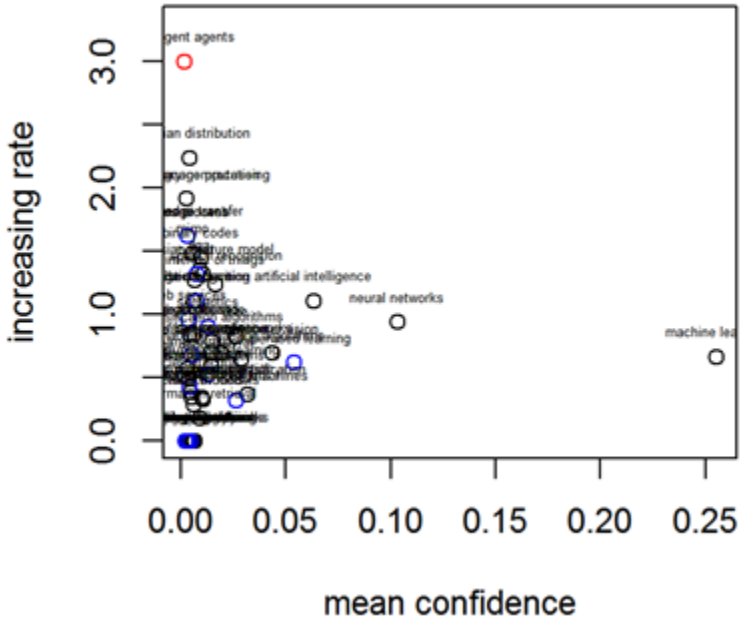
인공지능 분야 “Context awareness” 결과 예시

문서 빈도 추이 시각화

- Magnet word별 검색된 문서의 수 추이

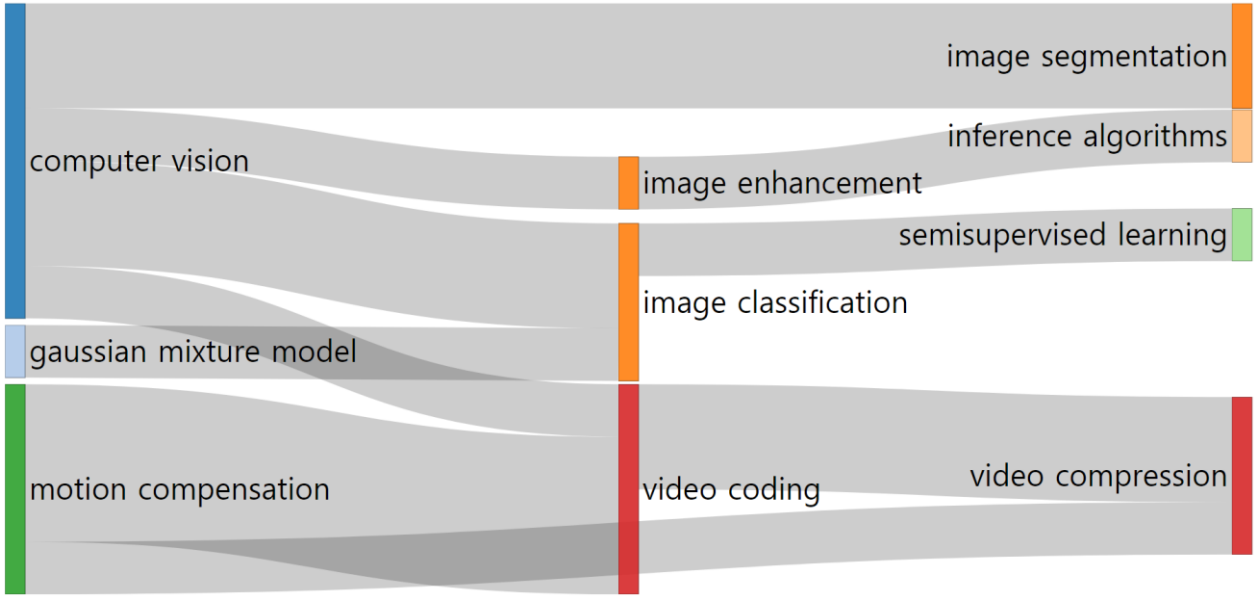


- 연도별 연관 키워드에 대한 신뢰도의 평균과 2017년 증가비율 산점도



문서 간 네트워크 시각화

- 연관 키워드 간 상관성을 시각화



인공지능 분야의 “Computer vision” 키워드의 네트워크 그림 예시

〈분석 모형의 고도화〉

- 연관성 분석과 Graphical Lasso 모형의 목적을 모두 반영한 분석 모형 제안 (Temporal Graphical Lasso)
- t시점에서 위험함수 $L_t(\gamma_t, \lambda_1)$를 최소화하는 γ_t 추정

$$L_t(\gamma_t, \lambda_1) = \sum_{j=1}^{p} \sum_{i=1}^{n} \left(x_{ij}^t - \gamma_{j0}^t - \sum_{k:k\neq j} \gamma_{jk}^t x_{ik}^t \right)^2 + p_{\lambda_1}(\gamma_t)$$

- 이 때, 벌점 함수는 아래와 같음

$$\mathrm{p}_{\lambda_1}(\gamma_t) = \lambda_1 \sum_{j=1}^{p} \sum_{k:k\neq j} |\gamma_{jk}^t|$$

- 여기에 시간의 변화에 다른 벌점 함수를 추가하여 회귀 계수 계산

$$\sum_{t=1}^{T} L_t(\gamma_t, \lambda_1) + p_{\lambda_2}(\gamma_1, \cdots, \gamma_T)$$

- 이 때, 추가되는 벌점 함수는 아래와 같음

$$p_{\lambda_2}(\gamma_1, \cdots, \gamma_T) = \lambda_2 \sum_{t=1}^{T-1} \sum_{j=1}^{p} \sum_{k:k\neq j} |\gamma_{jk}^{t+1} - \gamma_{jk}^t|$$

〈문서 정보 추출 방법 개선〉

- 단어의 포함 관계를 고려한 Keyword stemming을 활용, 문서에서 단어 정보를 추출

level0 level1 level2 level3

Aerospace and electronic systems

Aerospace engineering

Aerospace biophysics

Aerospace electronics

Aerospace safety

Air safety

Aerospace simulation

Aerospace testing

Satellites

Artificial satellites

Earth Observing System

Low earth orbit satellites

Moon

Space stations

참고문헌

- Chang, S. B., Lai, K. K., Chang, S. M. (2009). Exploring technology diffusion and classification of business methods: Using the patent citation network. Technological Forecasting and Social Change, 76(1), 107-117.
- Fan, W., Bifet, A. (2013). Mining big data: current status, and forecast to the future. ACM sIGKDD Explorations Newsletter, 14(2), 1-5.
- Friedman, J., Hastie, T., Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3), 432-441.
- Ho, J. C., Saw, E. C., Lu, L. Y., Liu, J. S. (2014). Technological barriers and research trends in fuel cell technologies: A citation network analysis. Technological Forecasting and Social Change, 82, 66-79.
- Santo, M., Coelho, G. M., Santos, Filho, L. (2006). Text mining as a valuable tool in foresight exercises: A study on nanotechnology. Technological Forecasting and Social Change, 73(8), 1013-1027.
- Technology Futures Analysis Methods Working Group. (2004). Technology futures analysis: Toward integration of the field and new methods. Technological Forecasting and Social Change, 71(3), 287-303.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.
- Tseng, Y.H., Lin C. J., Lin, Y. I, Text mining techniques for patent analysis, Information Processing and Management 43 (2007) 1216–1247.
- Yoon, B., Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. The Journal of High Technology Management Research, 15(1), 37-50.
- Yoon, B., Phaal, R., Probert, D. (2008). Morphology analysis for technology roadmapping: application of text mining. R&d Management, 38(1), 51-68.
- Zhu, D., Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. Technological forecasting and social change, 69(5), 495-506.