

텍스트 마이닝을 이용한 암반공학분야 SCI논문의 주제어 분석

정용복, 박의섭*

Keyword Analysis of Two SCI Journals on Rock Engineering by using Text Mining

Yong-Bok Jung, Eui-Seob Park*

Abstract Text mining is one of the branches of data mining and is used to find any meaningful information from the large amount of text. In this study, we analyzed titles and keywords of two SCI journals on rock engineering by using text mining to find major research area, trend and associations of research fields. Visualization of the results was also included for the intuitive understanding of the results. Two journals showed similar research fields but different patterns in the associations among research fields. IJRMMS showed simple network, that is one big group based on the keyword 'rock' with a few small groups. On the other hand, RMRE showed a complex network among various medium groups. Trend analysis by clustering and linear regression of keyword - year frequency matrix provided that most of the keywords increased in number as time goes by except a few descending keywords.

Key words Text mining, Data mining, Association rules, Clustering, Rock engineering

초 록 텍스트 형태의 자료에서 유용한 정보를 추출하는 텍스트 마이닝 기법은 데이터 마이닝의 한 분야이다. 본 연구에서는 암반공학 분야의 대표적인 국제 학술지인 IJRMMS와 RMRE에 2001년 이후 게재된 논문의 제목과 주요어를 대상으로 텍스트 마이닝 기법을 적용하여 주요 연구 동향과 시계열 트렌드, 연구 분야 상관관계 등을 파악하였으며 이를 이해하기 쉽도록 가시화하였다. 분석 결과 주요 연구 분야는 두 학술지 모두 유사하였으나 연관관계 분석 결과 IJRMMS의 경우 'rock'을 기반으로 1개의 큰 그룹과 소규모 그룹이 형성된 반면 RMRE는 중규모의 그룹이 형성되고 이 그룹 간에 연결이 형성되는 구조가 나타났다. 또한 시계열 자료로 변환하여 군집 분석과 각 주제어의 기율기 자료로 분석한 결과 일부 하강 주제어들이 있었으나 양적인 측면에서 차이가 있을 뿐 대부분 논문 수가 증가하는 것으로 나타났다.

핵심어 텍스트 마이닝, 데이터 마이닝, 연관관계법칙, 군집분석, 암반공학

1. 서 론

Google 검색으로 'mining'이란 키워드를 입력하면 관련검색어로 'data mining', 'text mining', 'opinion mining' 등의 키워드를 제시한다. 'mining'이 다른 분야에서는

광물이 아니라 빅 데이터(Big Data)에서 유용한 정보나 지식을 발굴하는 행위로 사용되고 있는 실정을 단적으로 보여주는 것이다. 데이터 마이닝이란 대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 찾아내는 것이며 다른 말로는 KDD (Knowledge-discovery in databases)라고도 한다(Wiki, 2015, Zhao, 2014, Pang-Ning et al., 2005). 데이터 마이닝이 이슈화된 배경에는 자료의 효율적 저장기술과 사물인터넷 기술의 발달로 방대한 양의 데이터 집적이 가능해졌으며 이런 빅 데이터의 분석을 통해 새로운 지식(유전자 정보, 고객 정보 등)의 발견이 가능해졌고 컴퓨터 성능의 향상으로 빅 데이터에 대한 실시간 분석이 가능해졌다는 점들이 있다. 현재 데이터 마이닝은 신용 평점

Received: Aug. 12, 2015

Revised: Aug. 26, 2015

Accepted: Aug. 27, 2015

***Corresponding Author:** Eui-Seob Park
Tel) +82428683098, Fax) +82428683416
E-Mail) espark@kigam.re.kr

KIGAM, 124 Gwahang-no, Yuseong-gu, Daejeon, 305-350, Korea

시스템, 사기탐지 시스템, 장바구니 분석, 최적 포트폴리오 구성 등 다양한 분야에서 사용되고 있다.

주로 숫자로 된 정형화된 데이터 분석에서 최근에는 텍스트, 이미지, 동영상, 음성 등 구조화되지 않은 비정형 데이터에 관심을 기울이고 있으며 이 중 인터넷 사용자의 증가에 따라 텍스트 마이닝(text mining)의 중요성이 부각되고 있다(Cho & Kim, 2012). 텍스트 마이닝은 데이터 마이닝의 한 분야이며 분석 대상이 디지털화된 텍스트이며 이로부터 유용한 정보를 추출하고자 하는 것이다.

Lee et al.(2008)은 특허문서로부터 기술동향 패턴을 파악하는 연구를 수행하였으며 특정 제품과 기술에 관한 특허의 진화패턴을 발견하여 차세대 개발 분야의 방향을 제시하였다. 조와 김(2012)은 산업공학분야 국제 학술지 논문에 대하여 텍스트 마이닝을 수행하여 주제어 빈도와 상관관계, 시계열 트렌드 등을 파악하였다. 이 외에도 경제동향분석(Song et al., 2013), 건설분야 트렌드 분석(Jeong & Kim, 2012), 신문사 논조 분석(Kam & Song, 2012) 등에 텍스트 마이닝이 활용되고 있다.

본 연구에서는 암석역학과 암반공학 관련 실험, 해석, 현장 적용 등의 연구 개발을 주도적으로 다루는 대표적인 국제저널인 International Journal of Rock Mechanics and Mining Sciences(이하 IJRMMS)와 Rock Mechanics and Rock Engineering(이하 RMRE)에 2001년 이후 2014년까지 게재된 논문의 제목과 주요어를 텍스트 마이닝 기법으로 분석하여 주요 연구 동향과 시계열 트렌드, 연구 분야 상관관계 등을 파악하였으며 이를 이해하기 쉽도록 가시화하였다.

2. 분석 대상

본 연구에서 사용한 데이터는 SCOPUS에서 제공한 자료를 사용하였으며 SCOPUS에서도 검색한 자료에 대해서 주요 저자, 국가, 소속 기관 등에 대한 정보를 그래프 형태로도 제공한다. 분석 대상은 2001~2014년 IJRMMS에 게재된 1,791편의 논문과 2002~2014년에 RMRE에 게재된 619편의 논문의 제목과 주요어이다. 2015년도 자료는 전체가 아직 게재된 것이 아니라서 분석에서 제외하였다. 제목과 주제어를 합쳐서 분석하였기 때문에 동일한 단어가 하나의 논문에서 2회 이상 나타날 수 있다. 연도별 패턴을 파악하기 위해서 발행연도도 분석에 사용하였다. SCOPUS에서 검색한 자료는 CSV 포맷의 파일로 내보내기가 가능하며 분석에는 CSV파일을 사용하였다.

3. 분석 방법

3.1 분석도구 R

분석에 사용한 프로그램은 통계 기반의 오픈소스 프로그램인 R을 사용하였다. R은 1995년 뉴질랜드의 오클랜드 대학의 Robert Gentleman과 Ross Ihaka에 의해 개발된 데이터 분석 프로그램이며 현재는 R Development Core Team에 의해 지속적으로 개선이 진행되고 있다. 비교적 사용하기 쉬운 환경과 무료, 다양한 플랫폼 지원 등 많은 장점을 가지고 있기 때문에 현재 데이터 분석 분야에서 Python과 더불어 양대 산맥을 구축하고 있다. R은 전통적인 통계분야 뿐만 아니라 개방성을 통해 다양한 패키지들이 제공되며 바이오 및 생물학 분야에서는 표준분석도구로 사용되고 있다(Yang & Kim, 2011). 본 연구에서는 R의 기본(base) 패키지와 tm(text mining), arules(association rules) 패키지가 주로 사용되었으며 이를 가시화하는데 wordcloud, arulesViz 등의 패키지를 사용하였다.

3.2 텍스트 마이닝

텍스트 마이닝을 위해서는 비정형의 대규모 텍스트 데이터를 효과적으로 처리하기 위한 자료구조가 필요하며 이에 대한 기본 플랫폼을 R에서는 tm 패키지가 제공한다. 가장 기본적인 데이터 구조는 말뭉치(Corpus)이며 이는 텍스트의 집합체이다. 2절에서 확보한 CSV파일을 읽어서 Corpus로 만들어야 이후 기본적인 텍스트 처리와 후속 분석 작업을 수행할 수 있다. Corpus가 구성되면 특정 주제어가 특정 문서(논문)에 몇 회 나타내는지를 표시한 문서-주제어 행렬을 구성할 수 있으며 이 행렬을 사용하여 주제어 빈도나 주제어별 연관성 등을 파악할 수 있다.

3.3 트렌드

연도별 특정 주제어의 변화를 파악하기 위해서는 위의 기본 자료를 연도별로 부분합을 구해야 한다. 이를 구해서 주제어-발간연도 행렬을 구성하여 기본적인 트렌드를 분석할 수 있다. 이 외에도 주제어-발간연도 행렬을 사용하여 군집분석을 수행하면 각 주제어를 트렌드가 유사한 그룹으로 분류할 수가 있다.

3.4 연관성 분석

연관성 분석은 데이터 안에 존재하는 항목 간의 연관 규칙(association rule)을 발견하는 것이다. 마케팅 분야에서 손님의 장바구니에 들어있는 품목 간의 관계를 알아본다는 의미에서 장바구니 분석(market basket analysis)이라고도 한다. 연관성 분석의 경우 transaction이라는 구조가 기본적으로 사용되며 이는 데이터베이스의 레

코드와 유사하다. 논문 자료의 경우 각 논문 제목과 주요어에 포함된 단어가 아이템이 되며 transaction은 이 단어(주제어) 중 어떤 단어가 해당 논문에 포함되었는가를 나타내게 된다.

논문자료에 대한 연관성 분석을 통해 연관 규칙을 찾을 수 있으며 연관성을 측정하는 기준으로는 지지도(support), 신뢰도(confidence), 그리고 향상도(lift)가 주로 사용된다. 파악된 연관규칙이 유용한 것이 되려면 일정 값 이상의 지지도와 신뢰도를 가져야 한다.

지지도는 전체 자료 중 두 항목(주제어 또는 상품)이 동시에 존재하는 자료의 비이며 다음과 같이 표시할 수 있다.

$$Support(A, B) = P(A \cap B) = \frac{A와 B가 동시에 포함된 논문}{전체논문} \quad (1)$$

신뢰도는 ‘If A, then B’와 같은 규칙에 해당하는 것으로 분모가 A를 포함하는 것으로 제한된다.

$$Confidence(A \rightarrow B) = \frac{P(A \cap B)}{P(A)} = \frac{A와 B가 동시에 포함된 논문}{A가 포함된 논문} \quad (2)$$

‘If A, then B’와 같은 규칙에 대해서 신뢰도가 높더라도 분모가 작은 경우 유용한 규칙이 될 수 없다. 이를 고려하기 위한 기준이 신뢰도이며 다음과 같이 정의된다.

$$Lift(A \rightarrow B) = \frac{P(A \cap B)}{P(A)P(B)} = \frac{P(BA)}{P(B)} = \frac{Confidence}{P(B)} \quad (3)$$

A와 B가 관련이 없다면 $P(B|A) = P(B)$ 가 되어 향상도는 1이 된다. 어떤 규칙의 향상도가 1보다 크다면 이 규칙은 결과를 예측하는 데 있어서 우연한 경우보다 우수함을 의미한다.

3.5 분석절차

텍스트 마이닝을 수행하는 일반적인 절차는 Fig. 1과 같다. 먼저 분석을 수행할 자료를 수집하고 해당 자료를 이용하여 말뭉치를 작성한다. 말뭉치가 작성되면 내장된 기능을 적용하여 기본적인 텍스트 처리(공백 제거, stop word 제거, 소문자로 변환 등)를 수행한다. 말뭉치로부터 문서-주제어 행렬을 생성하고 주제어를 단어단위로 추출한다. 주제어 중 단수와 복수로 표현된 단어

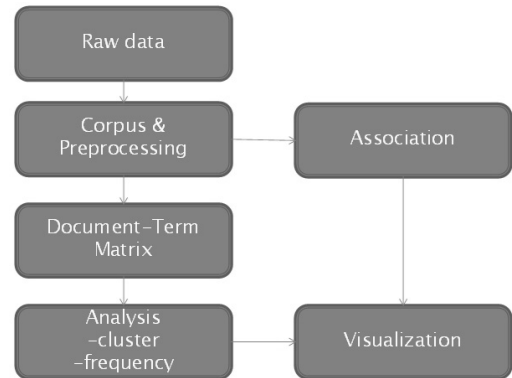


Fig. 1. Basic text mining procedure for the journal articles

가 있을 수 있으므로 주제어를 기반으로 전체 말뭉치에 대하여 복수로 표현된 단어를 단수로 변환한다. 분석에서 ‘study’와 같은 단어는 분석에 의미가 없을 것으로 판단하여 추가적으로 말뭉치에서 제거하였다. 수정된 말뭉치를 기반으로 다시 문서-주제어 행렬을 구성하여 주제어 빈도 등을 분석하고 이를 히스토그램이나 wordcloud 등으로 표현한다. 추가적으로 문서-주제어 행렬에 대해서 군집분석을 수행하고 이를 덴드로그램으로 표시하였다.

다음으로 주제어-연도 행렬을 구성하여 주제어의 연도별 변화 패턴을 파악하였다. 여기에서도 군집분석을 수행하여 연도별 트렌드가 유사한 그룹을 분류하고 표시하였다. 또한 각 주제어의 연도별 변화자료에 대해 선형회귀분석을 수행하여 구한 기울기를 근거로 주요 주제어를 분류하고 표시하였다. 마지막으로 연관분석을 수행하여 연관규칙을 추출하고 이를 가시화하였다.

4. 분석 결과

4.1 주요 주제어

수집한 논문 자료에 대해서 관사나 숫자 등을 제거한 상태에서 IJRMMS는 총 3,546개, RMRE는 2,083개의 단어가 각각 추출되었다. 이 단어 중 빈도를 기준으로 정렬하여 상위 20개의 단어를 Table 1에 빈도수와 함께 표시하였다. 두 학술지 모두 같은 암반공학에 대해서 다루기 때문에 단어나 순위가 아주 유사하게 나타났으며 ‘rock’이 압도적으로 많은 것을 알 수 있다. 20위 안에서 ‘stability’, ‘coal’, ‘element’, ‘deformation’, ‘mine’은 IJRMMS에만 나타났고 ‘joint’, ‘dynamic’, ‘slope’, ‘pressure’, ‘damage’은 RMRE에만 나타났다. IJRMMS의 경우 광산에 대한 주제도 명시적으로 다루기 때문에

Table 1. Most frequent keywords and their frequencies of each journal

Ranking	IJRMMS		RMRE	
	Keyword	Frequency	Keyword	Frequency
1	rock	1390	rock	649
2	stress	381	strength	180
3	fracture	380	fracture	172
4	analysis	346	analysis	158
5	numerical	310	tunnel	132
6	strength	295	test	127
7	mass	294	method	117
8	test	294	stress	112
9	model	284	model	105
10	method	276	numerical	104
11	modeling	261	failure	101
12	failure	260	mass	100
13	tunnel	188	modeling	100
14	effect	179	joint	96
15	stability	178	dynamic	78
16	coal	170	shear	73
17	shear	162	slope	65
18	element	154	pressure	64
19	deformation	153	damage	63
20	mine	152	effect	62

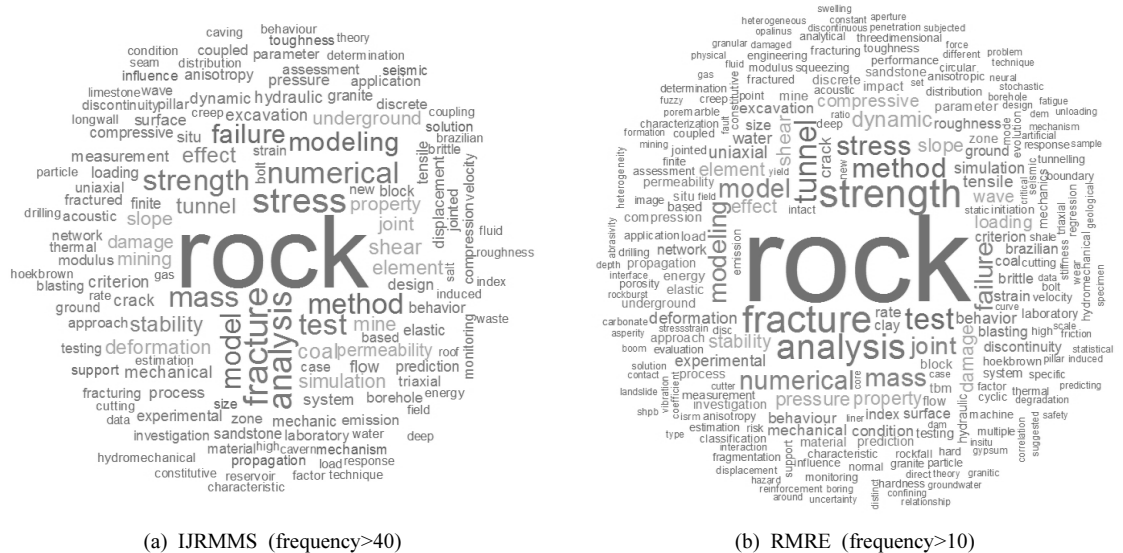
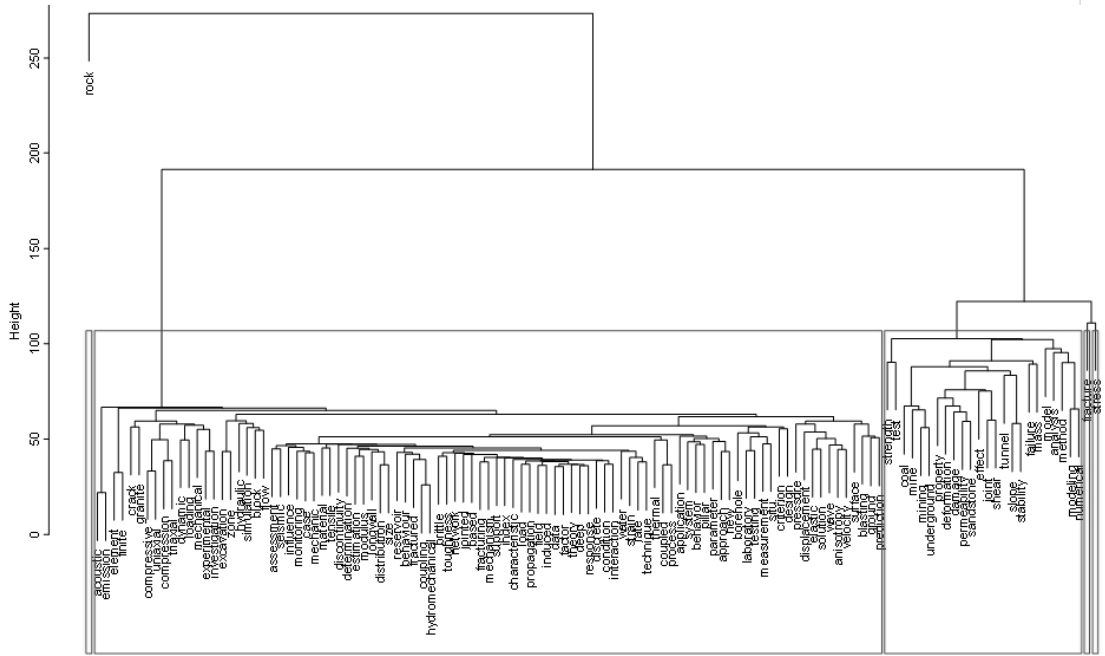


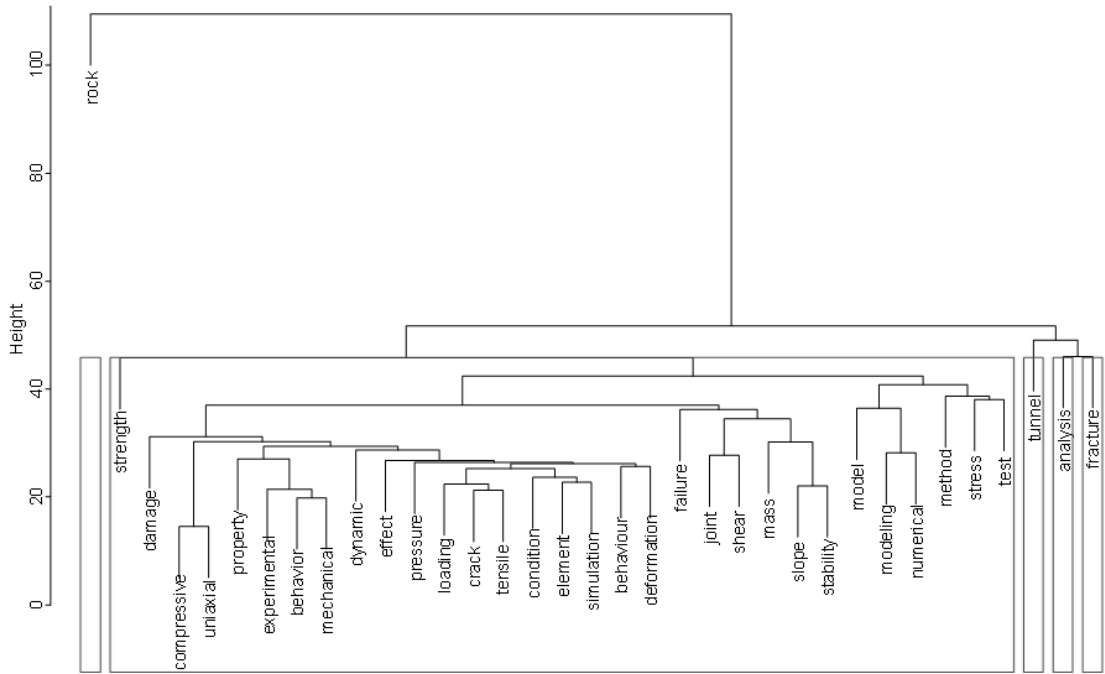
Fig. 2. Wordcloud of mined journal data based on the frequency of each keyword

‘coal’이나 ‘mine’이 상위권에 포함된 것으로 판단된다. 공학 분야에서는 위와 같은 자료는 일반적으로 히스토그램으로 표시한다. 텍스트 마이닝 분야에서는 히스토그램도 사용하지만 Fig. 2와 같은 wordcloud로 표시

하는 경우가 많다. 이 경우 빈도수에 대한 구체적인 정보는 표시되지 않지만 시각적으로 주요 주제어에 대해서 보다 직관적으로 파악할 수 있다. 문서-주제어 행렬에 대해서 유사도 또는 거리 기준으



(a) IJRMS



(b) RMRE

Fig. 3. Dendrogram of keywords with 5 classified groups

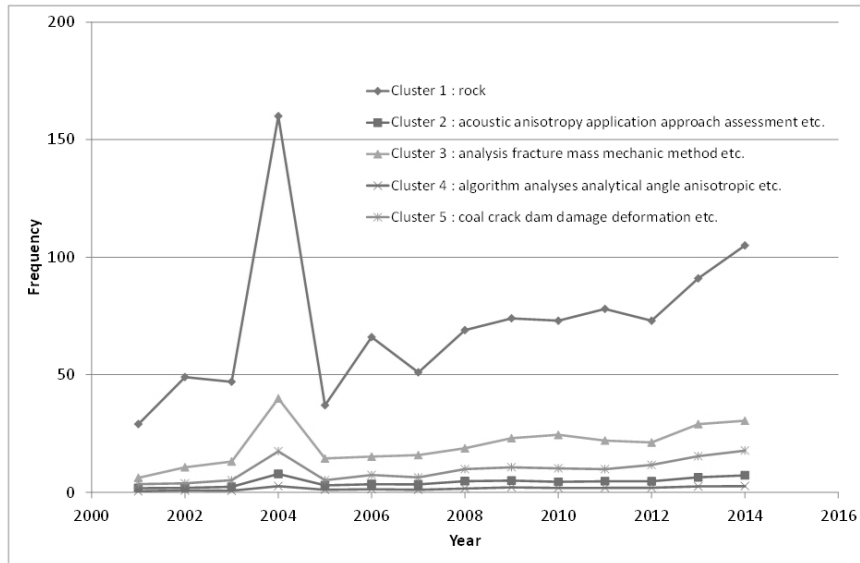


Fig. 4. Trend curves of 5 clusters of keyword from IJRMMS

Table 2. Clusters of term-year frequency matrix of IJRMMS

Cluster No.	Slope	Keywords
1	2.927	rock
2	0.292	acoustic, anisotropy, application, approach, assessment, based, behavior, blast, blasting, block, bolt, borehole, brittle, case, characteristic, compression, compressive, condition, coupled, etc.
3	1.191	analysis, fracture, mass, mechanic, method, mode, model, modeling, numerical, stress, test
4	0.132	algorithm, analyses, analytical, angle, anisotropic, aperture, area, around, artificial, back, band, bearing, boundary, brazilian, breakout, capacity, carbonate, cavern, caving, etc.
5	0.784	coal, crack, dam, damage, deformation, disc, effect, elastic, element, experiment, face, failure, formation, ground, joint, load, mechanical, mine, mining, property, ratio, shear, simulation, situ, stability, stone, strength, tunnel, underground

로 분류하고 이를 5개의 그룹으로 나누어서 표시하면 Fig. 3과 같다. IJRMMS은 ‘rock’ 그룹과 ‘stress’, ‘fracture’ 2개의 중간규모 그룹, 그리고 2개의 소규모 그룹으로 분류된 반면, RMRE의 경우 ‘rock’ 그룹과 이보다 작은 ‘tunnel’, ‘analysis’, ‘fracture’ 3개 그룹, 그리고 나머지 1개로 구분되어 IJRMMS의 경우 RMRE보다 ‘rock’ 하나의 주제어에 더 편중된 것으로 볼 수 있다.

4.2 주제어 트렌드

4.2.1 군집분석

주제어의 연도별 변화를 검토하기 위해 주제어-연도별 빈도 행렬을 구성하고 이 행렬에 대해 연도별 패턴을 기준으로 군집분석(k-means)을 수행하였다. 행렬 구

성 시 10회 이상 출현한 주제어로 한정하였으며 IJRMMS 은 365개, RMRE의 경우 244개의 주제어가 선정되었다. 5개의 군으로 나누는 결과 군의 중심이 연도별로 Fig. 4와 같이 나타났다. 여기서 군 번호는 특별한 의미는 없다. 5개의 군에서 1번 군은 ‘rock’이 단독으로 포함되며 가장 빈도수도 크고 연도별 증가추세도 크게 나타났다. 2004년은 Sinorock 학술회의에 제출된 논문이 모두 IJRMMS지널에 등재되어 다른 연도에 비해 압도적으로 논문 수가 많아서 그래프 상에 돌출되어 나타난 것이다. 나머지 4개 군의 주제어도 대부분 빈도수와 증가추세에서 차이가 있지만 계속 빈도가 증가하는 것으로 나타났다. 각 군별 주제어의 군별 변화추세에 대한 선형회귀결과와 구한 기울기는 Table 2와 같다. ‘rock’의 경우 기울기가 2.927로 연간 3회 정도 증가한 것이며 나

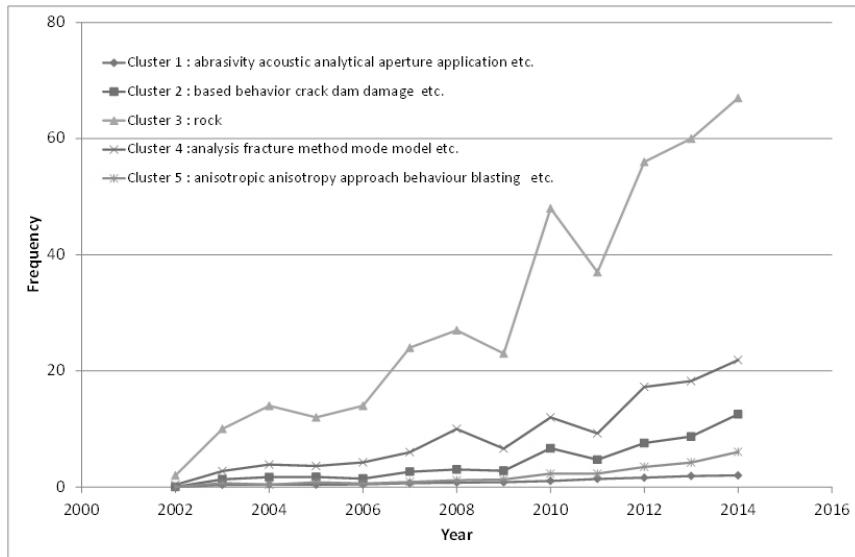


Fig. 5. Trend curves of 5 clusters of keyword from RMRE

Table 3. Clusters of term-year frequency matrix of RMRE

Cluster No.	Slope	Keywords
1	0.157	abrasivity, acoustic, analytical, aperture, application, around, artificial, asperity, assessment, block, bolt, boom, borehole, boring, boundary, carbonate, case, characterization, etc.
2	0.852	based, behavior, crack, dam, damage, deformation, disc, dynamic, effect, element, experimental, failure, formation, ground, joint, load, loading, mass, etc.
3	5.220	rock
4	1.610	analysis, fracture, method, mode, model, strength, stress, test
5	0.408	anisotropic, anisotropy, approach, behaviour, blasting, brazilian, brittle, characteristic, clay, coal, compression, compressive, condition, constitutive, coupled, criterion, deep, discontinuity, discrete, etc.

머지 군별 주제어들도 각각의 기술기만큼 평균적으로 증가한 것으로 볼 수 있다.

RMRE 저널 자료에 대해서도 동일하게 분석한 결과 Fig. 5와 같이 나타났으며 각 군별 기술기 및 주제어는 Table 3에 표시하였다.

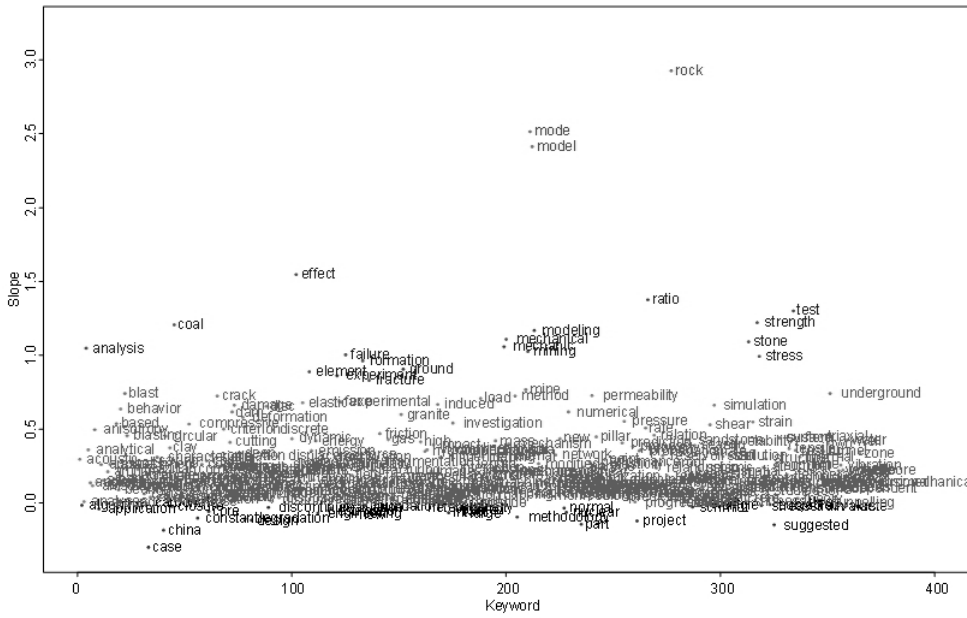
4.2.2 주제어별 트렌드

4.2.1절의 분석은 평균적인 트렌드를 분석한 것이다. 또 다른 분석 방법은 확보된 모든 주제어에 대해 연도 별 분포 패턴에 대해 선형회귀를 실시하여 기술기를 구하고 이 기술기를 근거로 상승, 유지 및 하강 주제어를 구분하는 것이다. 두 학술지 자료에 대해서 이 방법을 적용하고 기술기 값으로 표시한 것이 Fig. 6이며 Fig. 7은 기술기가 0 ~ 2 사이에 분포하는 주제어를 표시하

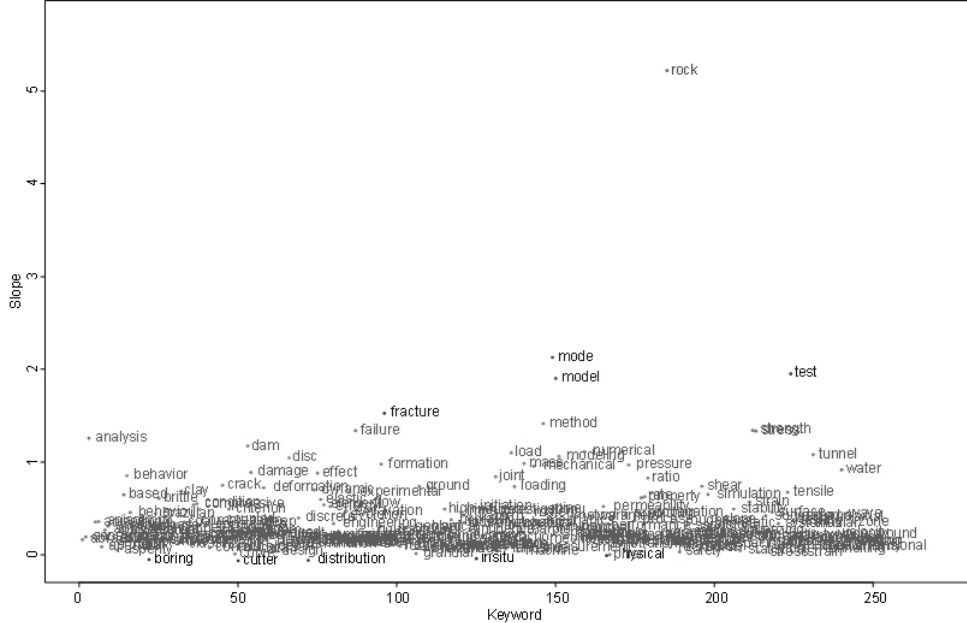
기 위해 y축 범위를 조절한 것이다.

두 학술지 모두 대부분의 주제어는 기술기가 0~1 사이에 분포하며 일부가 1보다 크거나 음의 값을 가지는 주제어도 일부 존재하였다. 본 연구에서는 기술기가 1 이상인 것을 상승, 0~1을 유지, 0보다 작은 경우 하강으로 구분하였으며 이를 Table 4에 표시하였다.

IJRMMS의 경우 RMRE에 비해 하강 주제어가 많이 나타났다. 그러나 ‘part’, ‘suggested’, ‘china’, ‘case’를 제외한 나머지 하강그룹의 주제어는 기술기가 거의 0에 가까운 것으로 연도에 따른 변화가 거의 없는 주제어들이다. RMRE의 경우 ‘physical’, ‘insitu’, ‘boring’, ‘distribution’, ‘cutter’가 하강 그룹에 속한다. 따라서 RMRE의 경우 모형실험과 TBM 분야가 다소 감소하고 있음을 알 수 있다.



(a) IJRMMS



(b) RMRE

Fig. 6. Distribution of slopes of whole keywords

4.3 주제어 연관관계

앞의 분석은 단순히 빈도나 추세를 파악한 것이며 이로부터 유용한 정보를 얻기는 다소 어렵다. 연관분석은

3절에서 언급한 바와 같이 각 주제어 간의 연관관계를 파악하여 현재 두 학술지가 어떻게 구성되어 있는지, 어느 분야가 중심을 차지하고 있는지 검토할 수 있다. 주로

Table 4. Classification of keywords based on the slopes

Group	Keywords	
	IJRMMS	RMRE
Rising (slope=>1)	rock, mode, model, effect, ratio, test, strength, coal, modeling, mechanical, stone, mechanic, analysis, mining, failure	rock, mode, test, model, fracture, method, strength, failure, stress, analysis, dam, numerical, load, tunnel, modeling, disc
Holding (0<=slope<1)	stress, formation, ground, element, experiment, fracture, mine, underground, blast, permeability, crack, method, load, experimental, face, elastic, induced, etc.	mass, formation, pressure, mechanical, water, damage, effect, behavior, joint, ratio, crack, shear, ground, loading, deformation, dynamic, clay, etc.
Declining (slope<0)	three, carbonate, shaft, subjected, single, algorithm, joint, foundation, stressstrain, value, closure, waste, schmidt, heterogeneity, discontinuity, evaluation, normal, application jointed, estimation, core, nuclear, intact, large, engineering, flow, methodology, constant, degradation, design, project, part, suggested, china, case	physical, insitu, boring, distribution, cutter

Table 5. Top 10 association rules of IJRMMS journals

No	Rules	Support	Confidence	Lift
1	{nuclear} => {waste}	0.013	0.920	51.491
2	{network} => {neural}	0.014	0.521	37.313
3	{emission} => {acoustic}	0.025	0.849	28.692
4	{coupling} => {hydromechanical}	0.011	0.500	22.962
5	{hoekbrown} => {criterion}	0.015	0.771	19.189
6	{anisotropic} => {anisotropy}	0.010	0.529	15.293
7	{uniaxial} => {compressive}	0.016	0.509	14.464
8	{fluid} => {flow}	0.011	0.594	14.370
9	{seam} => {coal}	0.013	0.852	13.997
10	{fracturing} => {hydraulic}	0.013	0.523	13.186

Table 6. Top 10 association rules of RMRE journals

No	Rules	Support	Confidence	Lift
1	{waste} => {nuclear}	0.011	0.778	68.778
2	{ism} => {suggested}	0.019	1.000	51.583
3	{machine} => {boring}	0.011	0.583	45.135
4	{acoustic} => {emission}	0.021	0.929	35.924
5	{opalinus} => {clay}	0.011	0.875	28.507
6	{neural} => {network}	0.015	1.000	26.913
7	{damaged} => {zone}	0.011	0.875	25.792
8	{specific} => {cutting}	0.011	0.636	24.619
9	{point} => {load}	0.018	0.846	22.773
10	{fluid} => {flow}	0.013	0.800	22.509

출되었다. 이 중 향상도 기준 상위 10개 규칙은 Table 5,6과 같다. 대부분 연관되어서 사용되는 단어들로 구성된 것을 알 수 있으며 RMRE의 경우 {opalinus}>

{clay}가 상위권에 있는 점이 특이하다. 이는 스위스에서 수행한 NAGRA 프로젝트의 일환으로 수행된 공학적 방벽 재료에 대한 연구결과가 다수 게재되었기 때문

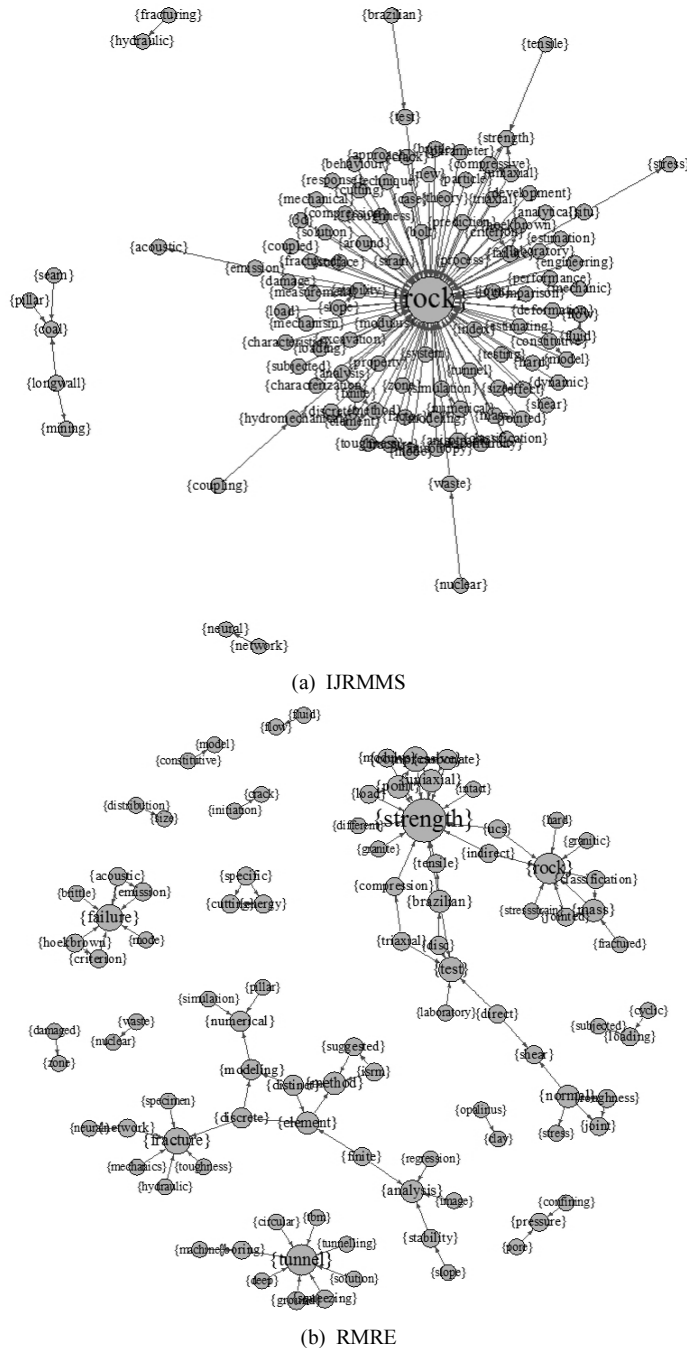


Fig. 8. Networks of association rules extracted from two journals (The size of text and symbol are proportional to the number of connected keywords)

이다. 이 외의 연관규칙들은 예상 가능한 것이라고 할 수 있다. 주제어 사이의 연관관계를 네트워크 그래프를 이용하

여 가시화한 결과는 Fig. 8과 같다. 글자와 원형 심벌의 크기는 다른 주제어와의 연결도 수에 비례하여 표시한 것이다. IJRMMS의 경우 'rock'을 중심으로 다양한 주

제가 고르게 연관되어 있으며 이외에 수압파쇄, 석탄광산, 신성망 등의 주제어 그룹이 별도로 나타났다. 반면 RMRE의 경우 서로 다양한 주제어 그룹을 형성하는 것으로 나타났다. 대표적으로 암석강도, 암반분류, 전단강도, 파괴역학, 터널링, 수치해석, 취성파괴 등의 큰 그룹과 유체유동, 구성모델, 균열, TBM, 공학적 차폐제, 공극압, 손상대, 폐기물 처분, 반복 하중 등의 다양한 소그룹이 형성되었다. 즉, IJRMMS는 암석과 암반의 특성과 석탄광산 분야가 주요 연구 주제이며 좀 더 근본적인 주제를 다루고 구조가 단순한 반면 RMRE는 암석과 암반에 대한 실험과 모델뿐만 아니라 터널, 수치해석 등 응용분야도 다루며 연관관계가 조금 더 다양하게 분포하였다.

네트워크로 연관관계 규칙을 표시하면 특정주제어 관련된 부주제들도 파악할 수 있다. RMRE에서 터널의 경우, ‘TBM’, ‘squeezing’, ‘ground’, ‘deep’, ‘solution’ 등의 주제어가 연관되어 있으므로 기계굴착과 심부터널, 그리고 이에 따른 지반조건에 관련된 논문들이 많다는 것을 알 수가 있다.

Fig. 8과 같은 연관관계 그래프는 지도도나 향상도와 같은 연관관계 지표 값은 표시되지 않아서 규칙의 특성을 파악하기 어렵다. 동일한 연관관계 규칙을 지도도와 향상도를 근거로 표시하면 Fig. 9와 같다. 여기서 원의 크기는 두 주제어 간의 지도도 값에 비례하며 원의 색이 짙을수록 향상도 값이 크다. 동일한 내용을 표시하지만 이 경우 특정 규칙이 더 강한 연관성이 있는지 쉽게 파악할 수 있다. 즉, Fig. 9(b)의 경우 {rock} => {mass}의 지도도가 가장 크고 {nuclear} => {waste}의 향상도가 가장 큰 것을 알 수 있다.

5. 토 의

5.1 국내 학술지와의 비교

국내의 암반공학 분야 전문 학술지인 ‘터널과 지하공간’에 대한 기존의 분석 결과와 비교할 때 ‘터널과 지하공간’ 학술지의 주제어 분포는 RMRE와 유사한 패턴을 보였다 (Jung, 2014). 즉, ‘터널과 지하공간’의 경우 ‘rock’, ‘tunnel’, ‘underground’, ‘analysis’의 4개 그룹과 ‘blasting’, ‘radioactive’, ‘mine’의 소규모 그룹으로 분류되어 3~4개의 주요 그룹을 중심으로 주제어들이 분포하여 ‘rock’ 그룹 1개 위주로 분포한 IJRMMS보다는 다양한 주제어 그룹으로 구성된 RMRE와 유사하였다.

단일 주제어에 대하여 2014년도 논문만을 대상으로 빈도수가 높은 주제어를 3개의 학술지에 대해 정리한 것이 Table 7이다. 3개 학술지의 상위권 주제어가 유사

한 것을 알 수 있으며 ‘ratio’가 상위권에 있는 점이 특이하다. 확인 결과 포아송비에 대한 것과 다양한 비율에 대한 연구들이 다수 게재되어 나타난 것이었다.

5.2 실험 vs. 수치해석

암반공학 분야는 연구 방법론 측면에서 볼 때 실험과 수치해석 방법으로 나눌 수 있으며 이들이 시대적으로 어떻게 변천했는지 파악하는 것은 추후 연구방향 설정에 있어서 의미가 있다. 게재된 논문은 실험 논문, 수치해석 논문, 그리고 이 둘이 다 포함된 논문이 있을 수 있으며 이를 기계적으로 구분하는 것은 아직 어렵다. 사전에 사람이 분류를 하고 이를 기반으로 학습을 시킨 뒤 예측하는 것을 시도해 볼 수 있으나 인력으로 처리할 자료의 수가 많아서 이는 추후 시도해 볼 만한 주제가 될 수 있다. 본 연구에서는 이 두 분야가 시대에 따라 어떻게 변천했는지를 알아보기 위해 단순히 특정 주제어 그룹을 정해서 비교해 보았다. 즉, 실험 그룹은 ‘experiment’, ‘experimental’, ‘test’, ‘laboratory’ 주제어 중 하나라도 포함한 경우, 수치해석 그룹은 ‘numerical’, ‘simulation’, ‘analytical’, ‘constitutive’ 주제어 중 하나라도 포함된 경우로 하였다. 이를 그래프로 Fig. 10~11에 표시하였다. 위의 기준으로 볼 때 현재까지는 실험 그룹이 수치해석 그룹에 비하여 양적인 측면에서 더 우위에 있음을 알 수 있다.

5.3 중국

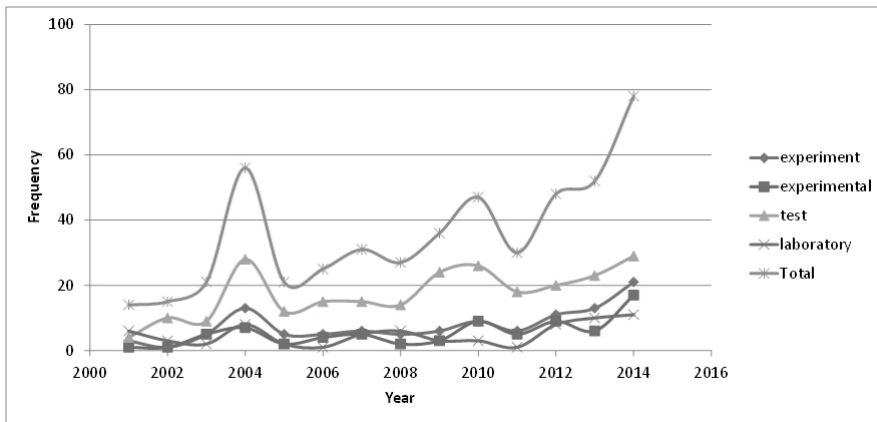
최근 암반공학 분야뿐만 아니라 대부분의 과학과 공학 분야에서도 중국이 다수의 논문들을 생산하고 있다. SCOPUS에서 제공하는 기본 분석도구를 사용하여 2001~2015년 까지 게재된 논문의 저자가 속한 국가에 대해 살펴본 결과는 Fig. 12와 같다. 중국의 논문 수가 압도적으로 많으며 China University of Mining Technology가 선두인 것으로 나타났다. 그다음으로 IJRMMS의 경우 Chinese Academy of Science와 Imperial college 순이며, RMRE의 경우 이탈리아의 Politecnico di Torino가 두 번째로 많은 논문을 산출한 기관으로 나타났다. 인구당 또는 해당분야 종사자 수를 기준으로 정규화하면 다른 결과가 나올 수도 있지만 현재 암반공학분야 논문에 있어서는 중국, 미국, 캐나다와 호주가 양적인 면에서 앞서 있음을 알 수 있다. 인용횟수를 근거로 추가 분석을 하면 질적인 측면에서 선두그룹을 형성하고 있는 국가나 기관을 추정할 수 있을 것으로 판단된다.

5.4 메타분석

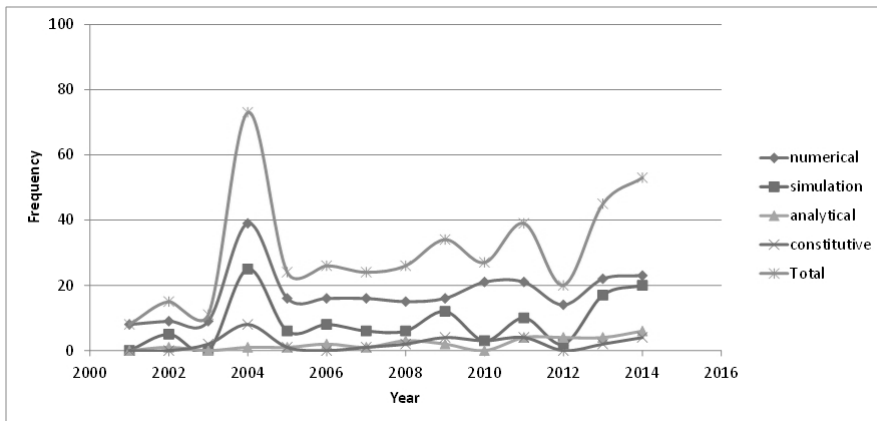
학술지 논문의 전문이 이미지 형태가 아니라 컴퓨터

Table 7. Top 10 keywords of 3 journals published in 2014

Ranking	KSRM	IJRMMS	RMRE
1	analysis	rock	rock
2	ratio	mode	mode
3	thermal	model	test
4	mine	stress	model
5	system	fracture	fracture
6	development	mechanic	method
7	based	test	dam
8	rock	ratio	failure
9	stability	effect	formation
10	case	analysis	damage



(a) Experimental group

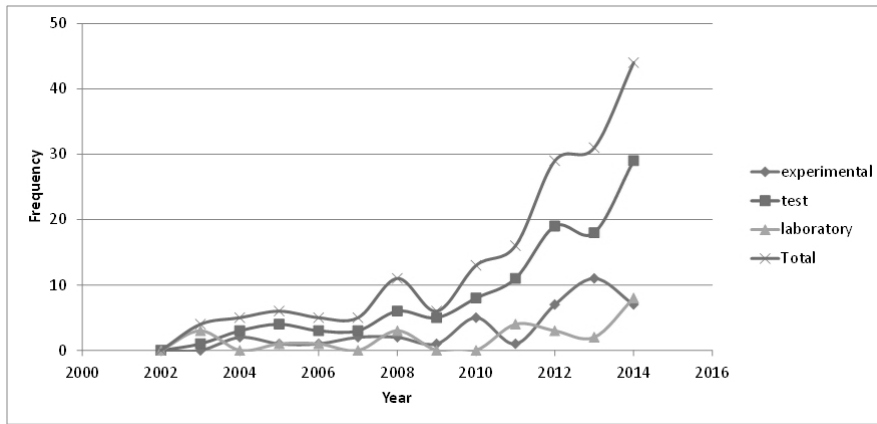


(b) Numerical group

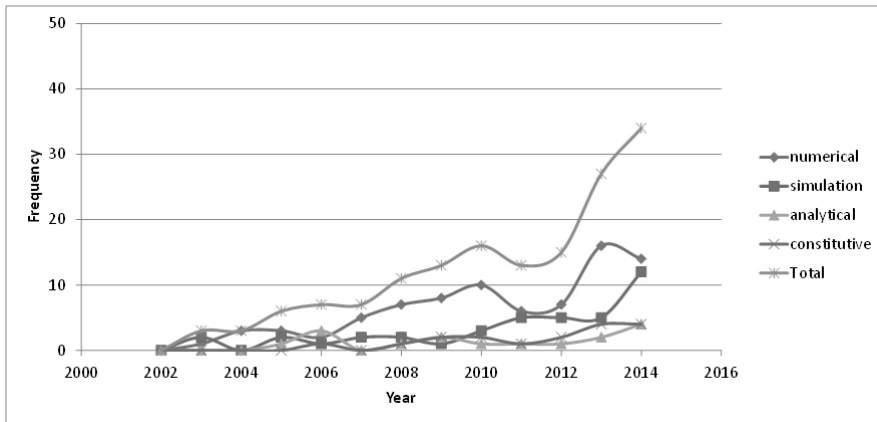
Fig. 10. Trend of two groups extracted from IJRMMS

가 이해할 수 있는 텍스트 형태로 제공될 경우 비교적 신속한 메타분석(meta analysis)이 가능하다. 메타분석

은 여러 연구 결과들을 수집하여 다시 분석하는 것이며 특정 분야나 주제에 관하여 누적된 연구 결과들을 종합



(a) Experimental group



(b) Numerical group

Fig. 11. Trend of two groups extracted from RMRE

하기 위해 통계적으로 분석하는 것을 말한다(Glass, 1976). 주로 사회과학과 의학 분야에서 많이 이루어지고 있다. 해당 분야의 경우 동일한 주제에 대하여 통계적으로 분석한 연구 결과들이 있을 경우 관련된 연구 논문들을 모두 수집하여 재분석을 하는 것이다. 암반공학 분야 논문에서 메타 분석을 수행할 수 있는 주제는 암석의 종류와 물리적, 역학적 특성 정보 분포, 수치해석에 사용한 구성 모델과 입력자료 분포, 특정 현상에 대한 영향인자 순위 도출, 지역/지질별 공학적 암반분류 분포 등이 있으며 앞으로 관련 논문들이 나오기를 기대해 본다.

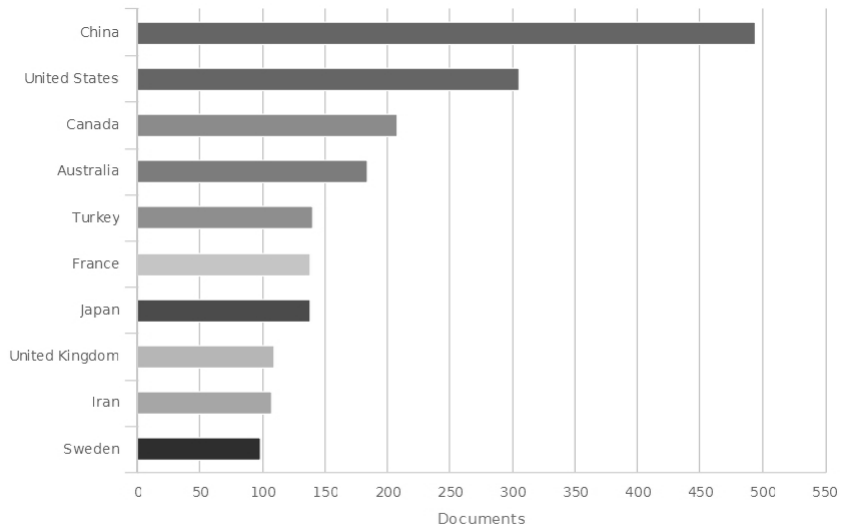
6. 결론

본 연구에서는 암반공학 분야의 대표적인 국제 학술지 IJRMMS와 RMRE에 2001년 이후 게재된 논문의

제목과 주제어를 텍스트 마이닝 기법으로 분석하여 주요 연구 분야와 시계열 트렌드, 연구 분야 상관관계 등을 파악하였으며 이를 이해하기 쉽도록 가시화하였다.

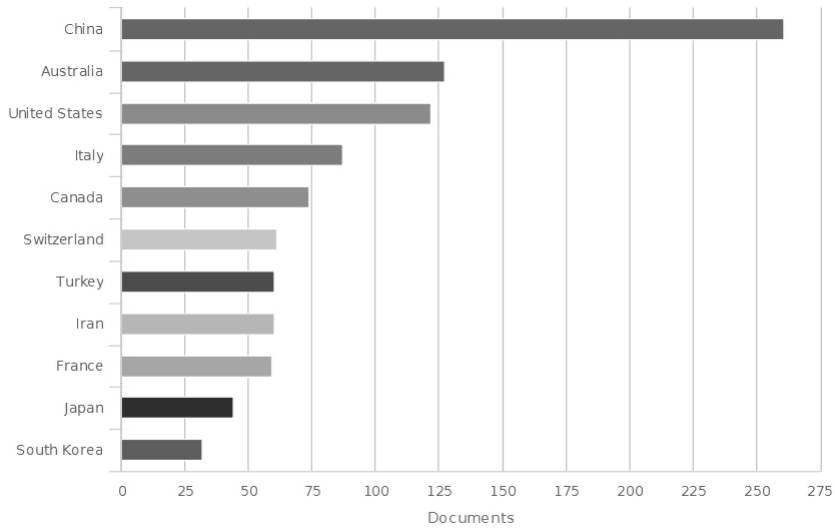
분석 결과 주요 연구 분야는 두 학술지 모두 유사하였으나 연관관계 분석 결과 IJRMMS의 경우 ‘rock’을 기반으로 1개의 큰 그룹과 소규모 그룹이 형성된 반면 RMRE는 중규모의 그룹이 형성되고 이 그룹 간에 연결이 형성되는 구조가 나타났다. 또한 시계열 자료로 변환하여 군집 분석과 각 주제어의 기울기 자료로 분석한 결과 일부 하강 주제어들이 있었으나 양적인 측면에서 차이가 있을 뿐 대부분 논문 수가 증가하는 것으로 나타났다.

끝으로 암반공학 분야에서 응용할 수 있는 논문 전문을 활용한 메타분석에 대해 간략히 살펴보았으며 앞으로 많은 관련 논문들이 게재되기를 기대한다.



Copyright © 2015 Elsevier B.V. All rights reserved. Scopus® is a registered trademark of Elsevier B.V.

(a) IJRMMS



Copyright © 2015 Elsevier B.V. All rights reserved. Scopus® is a registered trademark of Elsevier B.V.

(b) RMRE

Fig. 12. Country distributions of two journals (after www.scopus.com)

사 사

본 연구는 한국지질자원연구원의 주요연구사업인 ‘심부 지오시스템 특성 평가기술 개발(과제코드 GP2015-010)’의 일환으로 수행되었습니다.

References

1. Cho, S.G. and Kim S.B., 2012, Finding meaningful pattern of key words in IIE transactions using text mining, Journal of the Korean Institute of Industrial Engineers 38(1), 67-73.
2. Glass, G.V., 1976, Primary, Secondary, and Meta-Analysis

- of Research, Educational Researcher 5(10), 3-8.
3. Jeong, C.W. and Kim J.J., 2012, Analysis of trend in construction using textmining method, Journal of the Korean Digital Architecture Interior Association 12(2), 53-60.
 4. Jung, Y.B., 2014, Keyword analysis of KSRM journal by using text mining and association rules, Annual Conf. & General Assembly of Korean Society for Rock Mechanics (2014.11), 144-145.
 5. Kam, M.A. and Song, M., 2012, A Study on differences of contents and tones of arguments among newspapers using text mining analysis, Journal of Intelligence and Information System 18(3), 53-77.
 6. Lee, S., Lee, S., Seol, H., and Park, Y., 2008, Using patent information for designing new product and technology : keyword based technology roadmapping, R&D Management 38, 169-188.
 7. Pang-Ning T., Michael S. and Vipin K., 2005, Introduction to Data Mining (First Edition), Addison-Wesley Longman Publishing Co. Inc., Boston, 769p.
 8. Song, H.J., Park, K.S., Jung H.E. and Song, M., 2013, Trend analysis of Korean economy in the economic Literature by text mining techniques, Proc. of 20th Conf. of Korea Society for Information Management, 47-50.
 9. Wiki, 2015, https://ko.wikipedia.org/wiki/데이터_마이닝.
 10. Yang, K.S. and Kim, M.K., 2011, Introduction to R for Basic Data Analysis, Hannarae Publishing Co., Seoul, 310p.
 11. Zhao, Y., 2014, R and Data Mining: Examples and Case Studies, <http://www.RDataMining.com>.

정용복



1994년 서울대학교 공과대학 자원공학과
공학사
1996년 서울대학교 공과대학 자원공학
과 공학석사
2000년 서울대학교 공과대학 지구환경
시스템공학부 박사

Tel: 042-868-3097

E-mail: ybjung@kigam.re.kr

현재 한국지질자원연구원 지구환경연구
본부 심지층활용연구센터 지하공간연구
팀 책임연구원

박의섭



1989년 서울대학교 공과대학 자원공학과
공학사
1991년 서울대학교 대학원 자원공학과
공학석사
2000년 서울대학교 대학원 자원공학과
공학박사

Tel: 042-868-3098

E-mail: espark@kigam.re.kr

현재 한국지질자원연구원 지구환경연구
본부 책임연구원
