

# 검색 엔진을 이용한 키워드 연관성 분석

이주연\*, 노정현\*, 조수현\*, 이중화\*\*, 박유현\*\*  
동의대학교 컴퓨터소프트웨어공학과  
13969@deu.ac.kr

## The Keyword Relationship Analysis Using Searching Engine

Ju-Yeon Lee\*, Jung-Hyun No\*, So-Hyun Jo\*, Jung-Hwa Lee\*\*, Yoo-Hyun Park\*\*  
Dept of Computer Software Engineering, Dongeui University

### 요 약

대량으로 발생하는 키워드들 간의 연관성을 분석하고자 하는 연구는 꾸준히 진행되어 왔다. 많은 용어들의 관계를 분석하기 위한 방법으로 전문가 집단의 인력과 시간을 수행할 수 있지만, 시간과 비용이 많이 소모된다. 이를 해결하기 위한 방법으로 이미 관련 키워드 서비스를 제공하기 위한 시스템을 구축해 놓은 검색엔진을 사용해서 키워드들 간의 관계를 분석해 볼 수 있다. 본 논문에서는 IT분야의 논문에서 저자들이 자유롭게 작성하는 관심 분야를 키워드로 선정하고, 이 키워드들 간의 관계를 분석하기 위해 검색 엔진에서 출력하는 검색 결과 수를 사용한다. 검색 엔진에서 제공하는 검색 결과 수가 높을수록 다른 키워드와 연관성이 높은 키워드임을 알 수 있다.

### 1. 서론

최근 기술의 발달과 더불어 최신 용어들이 여러 분야에서 자주 등장하고 있다. 특히 IT 분야의 용어들은 짧은 시간 단위로 많은 용어들이 등장한다. 이 용어들을 적절한 위치에 분류하기 위해서 용어들의 연관성에 대한 분석이 필요하다. 용어들을 분류하기 위한 분류과정에서 전문가 집단을 통한 분석은 정확성을 높일 수 있지만, 현실적으로 많은 인력과 시간을 소모하게 된다.

본 논문에서는 IT분야의 논문에서 저자의 약력을 소개할 때, 자유롭게 기술되고 있는 저자의 관심분야를 키워드로 이용하여 검색 엔진에서 검색한 검색 결과 수를 통한 키워드 간의 연관성을 분석한다. 저자의 관심분야는 전통적인 키워드와 새로운 키워드가 함께 등장하기 때문에, 새로운 키워드와 전통적인 키워드와의 연관성 분석을 통해 새로운 키워드가 적합한 분야에 분류 되는데 참고 할 수 있다.

### 2. 관련연구

키워드들 간의 유사도를 측정하기 위한 선행 연구는 꾸준히 진행되어 왔다.

[1]에서는 학술 논문의 주제어를 중심으로 주제어 클러스터를 통하여 여러 연구 분야에서 사용한 주제어 및 그 사용빈도를 분석한다. 이 연구는 학술지에 수록된 논문을 주제어를 활용하여 주제적 관점의 계량적 접근을 하고 있기 때문에 기존의 연구 동향과 구별된다.

[2]에서는 잠재의미색인(LSI: Latent Semantic Indexing)

을 통해 핵심어를 포함하는 학술정보 데이터 키워드의 의미적 유사도를 계산한다. 여기서 잠재의미색인은 커다란 텍스트 코퍼스를 바탕으로 의미를 추출하고 표상한다. LSI는 공기 정보를 이용하여 같은 의미 공간에 분포하면 의미적으로 연관성이 있다고 판단하지만, 의미적 관련성이 적은 단어도 의미적으로 유사하다고 계산되는 경우가 발생하는 문제점이 있다.

[3]에서는 연관 클러스터링, 매트릭 클러스터링, 스칼라 클러스터링 3가지 방법을 이용해서 문헌 집합 내의 키워드 간의 유사도를 측정한다. 성능이 가장 좋은 매트릭 클러스터링에 많은 가중치를 두고 하나의 문헌에서 측정된 키워드 간의 유사도를 모두 더해 전체 문헌 집합에서의 키워드간의 유사도를 측정한다.

### 3. 연구방법

#### 3.1 자료수집



(그림1) 자료수집 순서도

본 논문에서는 키워드 간의 연관성을 분석하기 위해 IT 분야 논문에 기술된 저자의 관심분야를 이용한다. 다른 분야의 논문과는 다르게 IT 분야를 포함한 공학 분야 논문에서는 저자의 약력을 소개 할 때 관심분야를 함께 기술하기도 한다. 저자의 관심분야는 과거부터 사용되었던 용어와 새롭게 등장한 용어를 자유롭게 작성하기 때문에 다양한 용어가 등장한다. 그래서 저자가 작성한 관심분야를 키워드로 사용하게 되었다.

본 논문에서는 정보처리학회의 논문지인 KTCCS(KIPS Transactions on Computer and Communication System)에서 발간하는 컴퓨터 및 통신 시스템과 소프트웨어 및 데이터 공학 중 컴퓨터 및 통신 시스템 논문지로 한정하여 관심분야를 수집한다. 컴퓨터 및 통신 시스템 논문지의 2013년 60건, 2014년 47건의 논문을 이용하여 저자의 관심분야를 수집하였다.

관심분야 키워드 수집은 관심분야만을 따로 모아놓은 별도의 데이터베이스가 제공되지 않기 때문에 각각의 논문 파일을 열어 수작업으로 직접 논문 파일을 열어 관심분야를 별도의 파일에 입력해서 사용하였다.

먼저 총 107건의 논문 중 IT와 관련이 없는 키워드와, 정렬을 통해 중복 키워드를 제거하였다. 그리고 영어키워드와 한글키워드의 검색 결과 수의 차이가 발생하기 때문에, 한글 키워드로 대체 가능할 경우 주로 한글키워드를 이용하였다. 또한 ‘및’으로 구분된 키워드는 두 키워드로 나누어서 사용하였다. 이렇게 총 356개의 관심분야가 생성되었다.



(그림2) 필요로 하는 검색 결과 수

위의 그림처럼 본 논문에서는 (그림2)와 같이 구글에서 검색된 키워드의 검색 결과 수를 키워드들 간의 연관성을 분석하기 위해 필요로 한다. 정확한 결과를 얻기 위해서는 자료수집 단계에서 수집한 356개의 키워드를 모두 쌍을 만들어 검색을 해야 하지만 구글에서 대량의 검색을 할 경우 제한적인 검색을 제공하고 있다. 그래서 정확한 결과를 얻기 힘들고, 대량의 검색을 수행하기 때문에 검색을 수행하는 실험 시간이 오래 소요된다. 그렇기 때문에 전체적인 검색 횟수를 줄여 정확한 검색이 이루어 질 수 있도록 저자의 관심 분야를 수집하여 추출한 총 356개의 키워드를 구글에서 개별로 검색하여 검색 결과수가 제일 많은 상위 50개의 키워드를 본 연구의 키워드 간의 연관성을 분석하기 위한 키워드로 이용하였다.

키워드	검색수	키워드	검색수
인공지능	1,850	WSN	21.7
XML	741	FPGA	20.2
CAD	539	USN	19.9
SoC	219	ASIC	19.1
GIS	115	센서이동성	18.8
HTML5	109	내장형시스템	17.1
모바일	107	정보공개	16.6
클라우드컴퓨팅 보안	106	정보보호	16.2
임베디드 소프트웨어	88.7	로보틱스	16
U-HEALTHCARE	87.8	RFID	15.6
E-DISCOVERY	87	스마트센서	15.6
정보검색	65.2	PKI	13.9
실시간조합	58.8	MANET	12.6
스마트카드	54.7	유무선인터넷 멀티미디어통신	12
P2P	53.1	공개키암호	10.7
가상데스크톱	46.8	UI/UX	0.98
시뮬레이션	45.7	스마트폰	0.948
개인정보보호	44.7	RTOS	0.92
사이버보안	44.6	모바일오피스 서비스디자인	0.89
컴퓨터보안	36.7	부채널공격	0.83
NGN	36.5	모바일에이전트	0.56
빅데이터마이닝	32.8	그린ICT	0.69
VMS	32.2	프로그래밍언어	0.44
HADOOP	30.1	블루투스	0.33
바이오인포매틱스	24.3	얼굴인식	0.32

(그림2) 분석에 사용된 키워드 50개 (단위: 백만)

위의 표는 356개의 관심분야를 개별 검색하여 본 연구에서 사용할 상위 50개의 키워드와 검색 수를 나타낸 표이다. 검색 수는 백만 단위로 표시 하였으며, 개별 검색 수가 가장 많은 단어는 ‘인공지능’이며, ‘얼굴인식’의 검색 결과가 제일 적었다.

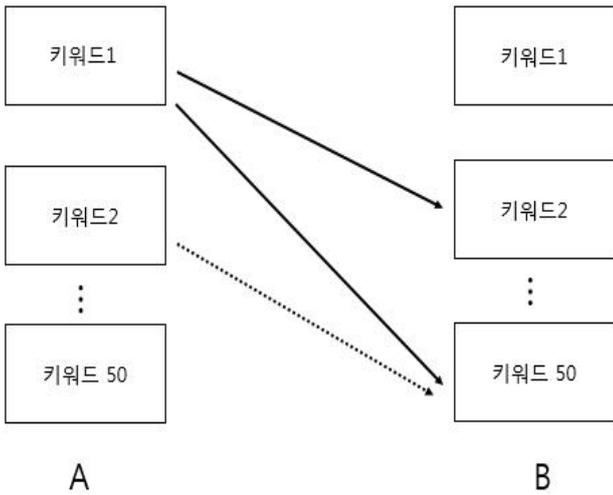
### 3.2 검색방법

수집한 키워드를 검색하여 활용하기 위해서는 검색 엔진에서 검색 한 후 검색 엔진이 제공하는 검색 결과 수를 필요로 한다. 본 논문에서는 검색 엔진으로 구글을 사용한다. 구글을 사용하는 이유는 구글을 포함한 4개의 검색엔진(네이버, 다음, 네이트, 구글)의 검색 결과를 비교해 봤을 때, 구글에서만 이 연구에서 가장 중요한 부분인 키워드의 검색 결과수를 출력해 주기 때문이다.

총 키워드 쌍의 개수는 다음과 같은 식으로 구해진다.

$$\text{키워드 쌍 총 개수} = n(n-1)$$

본 논문에서는 개별 검색결과가 상위 50개인 키워드로 키워드 쌍을 만들기 때문에 키워드 간의 연관성을 분석하기 위한 키워드 검색 쌍은 총 1225개의 조합이 생성되어 1225번의 검색을 행하게 된다.

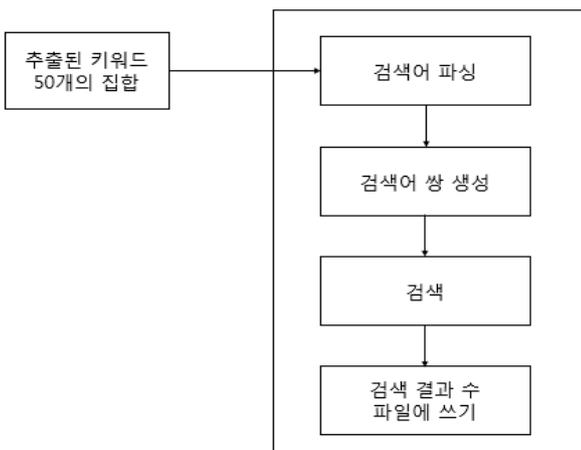


(그림3) 키워드 쌍 생성 방법

본 논문에서는 직접 1225개의 쌍을 구글에서 검색하는 방법 대신 PHP 스크립트를 이용하여 키워드 쌍을 생성하고, 검색 엔진 중 구글을 이용하여 검색하는 방법을 사용하였다.

검색은 최종적으로 추출한 50개의 키워드를 서로 중복되지 않는 쌍으로 만들어 이루어진다. A와 B에는 개별 검색 결과 수가 가장 많은 상위 50개의 단어를 각각 나열했다. 검색을 위해 중복되지 않는 쌍을 만들고(그림3) 이 쌍을 한 쌍씩 검색한다.

50개의 키워드 쌍의 조합인 1225회의 검색이 이루어지면 1225개의 검색 결과를 만든다. 그렇게 해서 최종적으로 작성되는 파일에는 구글에서 검색한 키워드의 쌍과 키워드 쌍의 검색 결과 수가 함께 작성된다. 그 후 검색 결과 수를 기준으로 검색어와 검색 결과 수를 내림차순으로 정렬한다.



(그림5) 전체 시스템 순서도

### 3.3 결과

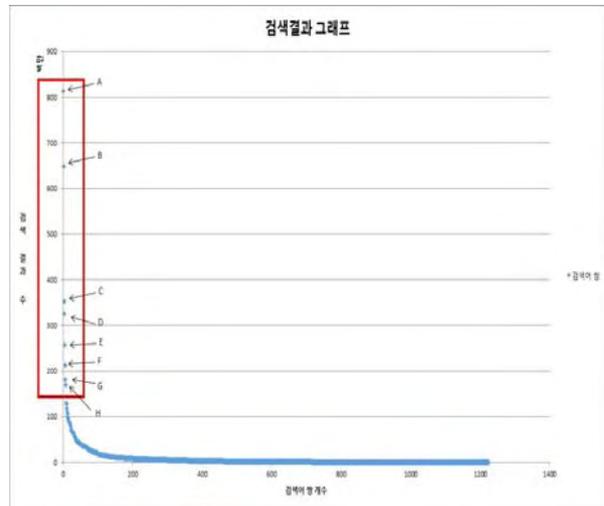
우선 50개의 검색어 중 어떤 키워드가 어느 계층에 많이 분포하는지 알아보기 위해 총 검색 결과 수 1225개를 상

위 계층에 검색 결과의 30%인 370개, 중간 계층에 40%인 485개, 하위 계층에 30%인 370개로 나눈다.

상위 계층에서 가장 많이 나타난 키워드는 'U-HEALTHCARE'가 32건, '시물레이션'이 29건, '센서이동성'이 27건으로 나타났으며, '클라우드컴퓨팅보안', '컴퓨터보안', '빅데이터마이닝', '얼굴인식'은 상위 계층에서는 한 건도 나타나지 않았다.

중간 계층에서는 '스마트센서'가 35건 '모바일에이전트'가 33건, '내장형시스템'이 26건으로 중간 계층에서 많이 분포했으며, 상위 계층에서 가장 많이 등장한 키워드인 'U-HEALTHCARE'는 7건, 'HTML5'가 5건의로 하위에 분포했다.

마지막 하위 계층에서는 개별 검색에서 중위권에 있었던 키워드인 '빅데이터마이닝'이 39건으로 제일 많이 등장하였고, 반면 'CAD'는 2건 밖에 검색되지 않았다.



(그림6) 검색결과 그래프

위의 그래프는 키워드 쌍들의 검색 결과 수를 엑셀 파일을 통해 내림차순으로 정렬 한 후 만든 키워드 쌍들의 검색 결과 수를 그린 그래프이다. 상위 8개의 검색 결과 쌍을 제외하고 대체적으로 키워드 쌍들 간 비슷한 검색 결과수를 가진다.

그래프상 표시되는 상위 8개 중 A는 제일 많은 검색 개수를 가지는 '인공지능/CAD'의 쌍으로 649,000,000건이 검색되었다. B는 '인공지능/U-HEALTHCARE'으로 353,000,000건 검색되었으며, C는 'U-HEALTHCARE/정보검색'쌍이 325,000,000건으로 검색되었다. 'XML/U-HEALTHCARE'쌍인 D는 257,000,000건으로 나타났으며, 그 뒤를 이은 검색 결과 수를 가지는 E는 'CAD/U-HEALTHCARE'로 213,000,000건이 검색되었다. F는 '시물레이션/센서이동성'의 쌍으로 검색 결과 수가 181,000,000건으로 나타났으며, 그 다음으로 표시된 G는 'U-HEALTHCARE/NGN'의 쌍으로 170,000,000건이 검색되었다. H는 'U-HEALTHCARE/FPGA'의 조합으로

130,000,000건이 검색되었다.

반면 하위권의 키워드 쌍을 살펴보면, ‘E-DISCOVERY/빅데이터마이닝’ 쌍이 3,460건, ‘부채널공격/바이오인포메틱스’ 쌍이 3,190건 ‘E-DISCOVERY/공개키암호’가 3,090건으로 검색 개수가 제일 적게 나타났다.

검색어1	검색어2	검색 결과 수
인공지능	CAD	649,000,000
인공지능	U-HEALTHCARE	353,000,000
U-HEALTHCARE	정보검색	325,000,000
XML	U-HEALTHCARE	257,000,000
CAD	U-HEALTHCARE	213,000,000
⋮		
컴퓨터보안	빅데이터마이닝	8,540
컴퓨터보안	WSN	3,850
E-DISCOVERY	빅데이터마이닝	3,460
부채널공격	바이오인포메틱스	3,190
E-DISCOVERY	공개키암호	3,090

(그림7) 검색 결과 수 상위 5개와 하위 5개

키워드 쌍의 두 키워드가 개별 검색 순위에서 높은 검색 결과 수를 가질수록 키워드 쌍도 높은 검색 결과 수를 가지게 되며, 개별 검색 결과 수가 낮은 키워드를 포함하고 있는 쌍은 비교적 낮은 검색 결과 수를 가진다.

개별 검색어의 검색 결과 수 순위에서 상위에서 10번째 검색 결과 수를 가진 ‘U-HEALTHCARE’가 키워드 쌍으로 검색 했을 경우에도 상위권의 검색 결과 수 순위에 위치해 있다. 그렇기 때문에 키워드 ‘U-HEALTHCARE’는 여러 키워드와 연관이 있는 키워드로 높은 검색 결과 수를 나타내는 키워드임을 알 수 있다.

#### 4. 문제점 및 향후 연구과제

본 논문에서 키워드 쌍의 검색을 수행하는 구글은 대량의 검색을 수행할 경우 제한적으로 검색을 제공하고 있다. 이와 같은 이유로 대량의 키워드 쌍을 검색해야 하는 현재 시스템에서는 다수의 키워드 쌍 모두가 검색을 수행하기에는 무리가 있다.

이 연구가 더 정확한 정보를 도출해 내기 위해서는 많은 키워드를 이용한 검색 결과 수를 필요로 한다. 하지만 현 시스템은 단순히 구글 URL로 검색을 행할 수밖에 없어 제한적인 정보를 제공하지만, 구글 API를 이용하여 키워드 쌍의 검색을 시도하거나, 한 대의 컴퓨터에서는 이루어지는 제한적인 검색 대신 대량의 키워드 검색이 가능하도록 클라우드 시스템을 이용한 분산 시스템을 통해 효율적으로 키워드 쌍을 검색하고, 검색에 소요되는 시간을 줄이고자 한다.

또한 검색 결과 수를 통한 키워드 간의 상위 관계에 대한 면밀한 분석을 필요로 하는데, 현재 키워드 간의 관계를 분석하기 위한 지표가 부족하기 때문에, 향후에 키워드

쌍 간의 검색 결과의 정확성을 높이고, 키워드 간의 관계를 분석하기 위한 방법으로 분류표를 활용하거나 다양한 방법을 이용하여 키워드 쌍 간의 면밀한 분석을 시도하고자 한다. 분류표는 기존에 검증되어 있는 분류표를 활용하거나 새로운 분류표를 구축하여 활용하는 것도 하나의 방법이 된다.

이 연구에서 구축된 시스템은 현재 정보처리학회의 논문지인 컴퓨터 및 통신 시스템의 107건의 논문 중 356개의 관심 분야를 추출 하여 개별 검색을 통한 검색 결과 수가 가장 높은 상위 50개의 키워드만을 제한적으로 사용하고 있다. 그렇게 때문에 차후에는 키워드를 수집 할 수 있는 전 분야의 논문을 대상으로 하여, 제한적인 키워드를 키워드 쌍을 만들어 검색하는 대신 대량의 키워드를 사용한 키워드 쌍의 결과 수를 수집하여 정확한 정보로 활용하고자 한다.

#### 5. 결론

본 논문에서는 검색 결과 수가 높은 키워드가 키워드 간의 유사도가 높거나, 다른 키워드와 깊은 관련이 있는 키워드 일 것이라고 가정하기 때문에, 검색 엔진에서 제공하는 검색 결과 수가 매우 중요한 정보가 된다. 그렇기 때문에 검색 엔진 중 검색 결과 수를 출력해 주는 구글을 사용하였다. 향후 다수의 검색 엔진에서 검색 결과 수를 얻을 수 있다면 더 정확한 키워드 간의 관계를 분석할 수 있을 것이다.

검색 엔진에서는 검색어와 유사한 키워드나, 관련 있는 키워드를 제공해 주는 서비스를 진행해 왔다. 물론 다양한 정보가 존재하는 검색 엔진에서 제공하는 높은 검색 결과 수에 전혀 관련이 없는 정보가 포함되어 있을 수도 있다. 하지만 두 키워드가 유사도가 높은 키워드 이거나, 두 키워드 간의 연관성이 높은 경우 더 많은 양의 검색 결과를 제공하게 된다.

현실적으로 인력과 시간이 소요된 전문가 집단이 정의하는 키워드 간의 관계는 명확할 것이다. 하지만 키워드 간의 관계를 정의할 때 이미 그에 대한 서비스를 진행해 왔던 검색 엔진에서 키워드를 검색한 검색 결과 수를 이용하는 것도 키워드 간의 관계를 정의하는 하나의 방법이 될 수 있다.

#### 참고문헌

[1] 이해영, 광승진, “국내 학술지 논문의 주제어를 통한 학술연구분야 관계분석”, 한국비블리아학회지, 제22권, 제3호, 2011.

[2] 조민희, 정도현, “학술정보데이터의 키워드 연관성 분석”, 한국인터넷정보학회지 추계학술발표대회 논문집, 제11권, 제2호, 2010.

[3] 이상훈, 김기태, “클러스터링 기법을 이용한 키워드 유사도 순위화 알고리즘에 따른 사용자 질의 확장”, 한국정보과학회 봄 학술발표논문집, 제30권, 제1호, 2003.