K-평균 군집화 기법을 활용한 DBLP 논문 서지정보의 연대별 출현 패턴 연구

허주성, 임현교, 김경한 한연희*
한국기술교육대학교 컴퓨터공학부
e-mail:{chil1207, glenn89, goslim56, yhhan}@koreatech.ac.kr

Finding Meaningful Chronological Pattern of Key Words in Computer Science Bibliography

Joo-Seong Heo, Youn-Hee Han*
School of Computer Science and Engineering
Korea University of Technology and Education

요 약

컴퓨터공학 분야의 논문 정보를 다루고 있는 대표적인 사이트인 DBLP의 연구 동향을 알아보기 위해 본 논문에서는 약 300만개 이상의 논문 서지정보 가져와 분석했다. IT용어 사전을 만들고 각 논문의 제목과 초록에 포함된 주제어를 추출해 분석을 위한 고차원의 행렬을 만들고, k-평균 군집화 기법을 활용하여 1960년도부터 2010년도까지 총 60여 년간의 연대별 주제어 출현 패턴을 분석함으로써 흥미로운 결과를 도출해 냈다.

1. 서론

데이터마이닝(Data Mining)은 방대하고 복잡한 데이터 내부에 존재하는 유용하고 의미 있는 정보를 이끌어 내는 방법을 연구 하는 학문이다. 주로 숫자 형태의 일정한 데이터 구조로 정형화된 데이터(structured data)를 분석해오던 데이터 마이닝의 연구자들은 최근 들어 텍스트, 이미지, 동영상, 음성 등과 같이 구조화되지 않은 비정형 데이터에 관심을 기울이고 있다. 특히 비정형 데이터 중에서 최근 인터넷 사용자의 폭발적인 증가에 힘입어, 웹 마이닝(Web Mining)과 텍스트 마이닝(Text Mining)의 중요성이 더욱 부각되고 있다. 이러한 관심은 대량의 텍스트에 대한 분석 및 연구로 이어져 생물학의 유전자 정보, 기술경영의 특허정보 등의 영역에서 적극적으로 연구되고 있다. 특히 특허문서로부터 기술 동향의 패턴을 알아보는 연구가 행해 졌다. 이 연구는 특허문서의 제목과 초록으로부터 주제어를 추출한 후, 주제어의 출현 빈도수로 기술 동향을 파악하고 주제어간의 관계를 파악해 기술과 제품의 상관관계에 대한 규명을 하였으며 나아가 특정 제품과 기술에 관한 특허의 진화패턴을 발견하여 차세대 개발 분야의 방향을 제시하였다. 이 외에도 특허문서를 텍스트 마이닝 기법을 이용해 분석한 연구가 행해졌다.

본 논문에서는 컴퓨터공학 분야의 논문 정보를 다루고 있는 대표적인 사이트인 DBLP로부터 약 300만개 이상의 논문 서지정보를 가지고 있는 xml 파일을 추출, 분석했다. 또한 추출한 논문 서지정보를 활용해 여러 저널 홈페이지로부터 유용한 정보를 이끌어내는 웹 컨텐츠 마이닝(Web Content Mining) 연구를 수행하였다. IEEE, Springer, ACM 등의 주요 저널 홈페이지에서 수집한 초록을 관찰하여 연대별 주제어의 출현횟수를 기록한 고차원의 데이터를 생성하고, 1940년대부터 2010년대까지 DBLP의 연구동향을 논문의 주제어의 k-평균 군집분석을 통해 살펴보았다.

* 교신 저자 : 한연희

본 과제(결과물)는 교육과학기술부의 재원으로 지원을 받아 수행된 산학협력 선도대학(LINC) 육성사업의 연구결과입니다.

2. 데이터

본 논문의 연구대상으로 선정된 DBLP는 1935년 이후 현재까지 컴퓨터공학 분야의 논문을 300만 개 이상 보유하고 있는 대표적인 사이트다.

본 연구에서는 DBLP에서 보유하고 있는 3251714 개의 모든 논문을 분석 대상으로 삼았다. 또한 관찰대상이 되는 주제어는 3천여 개의 IT용어 사전(Gartner IT Word)을 만들고 모든 논문에 기록된 주제어 중에서 불용어, 범용어를 제외한 IT용어 사전에 기록된 단어를 추출하여 분석했다. 선정된 주제어는 총 1397개로 그 중 일부는 다음과 같다. network, internet, web, openstack, nfv, deep learning, protocol, framework, tcp, bluetooth ..... data center, mobile ip, database desine, sna, ftp 수집한 항목과 이를 위해 활용된 프로그램의 요약은 <표1>과 <표2>에서 보여주고 있다.

<표1> DBLP 논문의 주제어 연구 대상 개수

구 분	IT용어사전	연구 대상
주 제 어	2896	1397

<표 2> 데이터 수집을 위한 항목

항 목	내 용
수집 항목	제목, 출판년도, 초록
수집 프로그램	Python, Java
데이터베이스	MySQL
분석 도구	Spark, Python

이와 같이 수집된 DBLP 논문의 초록에서 선정된 주제어의 출현빈도를 관찰하고, 1940년대부터 2010년대까지 8개의 연대로 나누어 주제어의 관측치를 기록한 1397×8 행렬을 생성하였다. 또한 <표 3>은 연대별 논문 개수 정보를 보여주고 있는데 대부분의 논문이 최근 20~30년간 존재하는 것을 확인할 수 있다. 한편, 8개의 연대로 나누어 행렬을 생성했을 때, 처음 2개의 연대는 논문의 개수가 적어 IT용어 출현빈도가 매우 적었다. 따라서 본 논문에서는 1940~50년대를 제외한 1960년대부터 2010년대까지 총 6개의 연대를 대상으로 한 분석을 통해 DBLP의 연구동향을 살펴보았다.

<표 3> 연대별 논문 개수 정보

구 분	1956~1965	1966~1975	1976~1985
논문 개수	5096	22827	66566

1986~1995	1996~2005	2006~2015
24819	795146	2113389

<표4>을 살펴보면, 주제어 'network'는 1960년대부터 2010년대까지 꾸준히 관찰됨을 알 수 있다. 하지만 주제어 'openstack'은 2010년대에만 160번 관찰되는 것을 확인 할 수 있다. 위 데이터는 다음 장에서 소개하는 k-평균 군집화 기법을 활용하여 분석하게 되는데 <표4>의 데이터는 주제어의 연대별 출현 패턴을 알아보는데 활용되었다.

<표4> 연대별 주제어 출현 빈도

	Y1960	Y1970	…	Y2010
network	69	217	…	151642
openstack	0	0	…	160
…	…	…	…	…
data center	0	2	…	4043

3. k-평균 군집화 기법

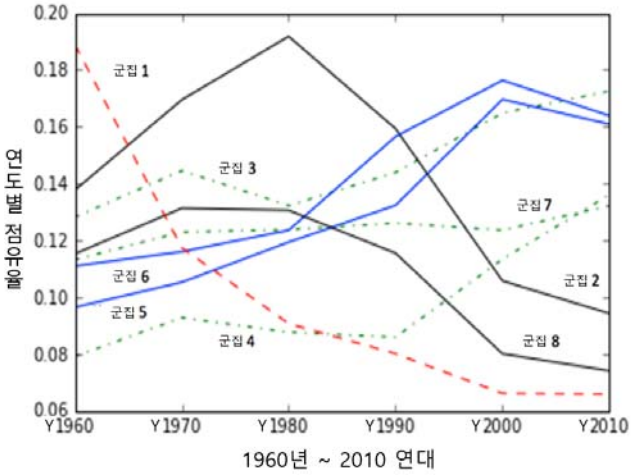
k-평균 군집화는 각 관측치의 최소평균거리를 이용하여, 전체 데이터를 k개의 군집으로 분할하는 군집기법의 하나이다. k-평균 군집화 기법을 간단히 요약하자면, 먼저 k개의 초기 점(seed point)을 임의로 선택하고, 각각의 관측 치와 k개의 초기 점과의 거리를 계산하여 해당 관측치를 가장 가까운 초기 점에 배정하는 k개의 군집을 형성한다. k개의 군집이 생성되면 각 군집의 평균점을 새로 계산하고 새로이 결정된 평균점과 각각의 관측 치와의 거리를 계산하여 새로운 군집을 형성한다. 위 단계를 반복하여 더 이상 평균점이 바뀌지 않아 군집형태가 동일하게 유지 될 때 최종 군집결과가 결정된다. k-평균 군집화 기법을 사용하기 위해서는 먼저 군집수(k)와 거리 척도를 결정해야 한다. 군집수를 결정하는 방법들은 몇몇 존재하고 있으나 어느 하나의 방식이 좋다고 말하기는 어려우며 보통 문제에 대한 배경지식을 기반으로 방법을 결정한다. 거리계산방식은 유클리드(Euclidean), 맨하튼(Manhattan), 상관관계(Correlation) 등 다양한 방식이 있으나 데이터의 특성과 분석 목적에 맞게 결정을 한다.

4. 실험 결과

본 장에서는 군집 분석으로 '논문 주제어의 연대별 출현 빈도 데이터'를 이용한 k-평균 군집화를 수행하였다. 이는 1960년대부터 2010년대까지의 각 주제어들의 연대별 출현 패턴을 알아가 보기 위함이다. 군집개수(k)는 5~15개로 나누어 분석해 보았고, 그 결과 8개로 분석 했을 때 가장 의미 있는 결

과를 얻을 수 있었다. 거리방식으로는 주제어 출현 빈도의 시계열 패턴을 반영하기 위해 상관관계거리를 이용하였다. 한편, k-평균 군집화에 앞서 연대별 논문 개수가 모두 다르기 때문에 정확한 결과 값을 얻기 위해 연대별로 주제어 출현 빈도에 대한 정규화를 시행했다.

(그림 1)은 군집화의 결과를 이용하여, 각 군집에 해당하는 주제어들의 평균 출현정도를 연대별로 보여주고 있다. 8개의 군집은 크게 네 가지 패턴으로 설명 할 수 있는데, 주제어 출현양의 추세가 최근 증가하고 있는 군집, 관찰 정도가 꾸준히 하강하고 있는 군집, 마지막 연대에 들어 급격히 하강하는 군집, 과거 특정 연대에서 급격히 증가했었던 군집 등으로 분류 할 수 있다.



(그림 1) 8개 군집의 연대별 주제어 출현 빈도

<표 5> 4가지의 그룹으로 분류한 주제어 군집 결과

구분	군집	주제어	비고
그룹 1	군집 3 군집 4 군집 7	access point, active directory, adaptive learning, affective computing, analytic applications, analytics, application architecture, authentication, technologies, autonomous system, autonomous vehicles, backbone, backbone network, bandwidth, base station, batch processing, benchmarking, bioinformatics, business analytics, content management, concurrent engineering, cognitive radio, clustering, cloud computing, circuit switching, checksum, code division multiple access, design thinking, deep learning, data scientist, data center, distributed data management, dynamic routing, eye tracking, femtocells, fraud detection, grid computing, html5, internet protocol, ipv6, job scheduling, media access control, mobile ip, mobile network, mobile social networks, multithreading, network intelligence, network management, network virtualization, openstack, p2p, protocol stack, proxy servers, saas, sdh/sonnet, session initiation protocol, social analytics, social network analysis, software defined networking, switched network, tcp/ip, text analytics, virtual reality, virtualization, vlan, vpn, web crawler, wireless broadband	최근 상승 주제어
그룹 2	군집 5 군집 6	architecture, availability, broadcast, buffer, cache, clock, community, data mining, encryption, framework, fuzzy logic, information technology, java, load balancing, machine learning, mobile, protocol, qos, scalability, signature, streaming, tcp, throughput, visualization, web, web services, xml	상승 & 최근 하강 주제어
그룹 3	군집 2 군집 8	artificial intelligence, attenuation, broadband, c++, channel capacity, circuit board, composition, cycle time, database design, database management system, distributed computing, distributed database, expert system, hypertext, information management, knowledge representation, microprocessor, operation system, packet switching, parallel processing, response time, software development, speech recognition, synchronization, terminal, transducer, workstations	한 때 상승 했었던 주제어
그룹 4	군집 1	binary code, modulation, redundancy, switch, control, application, information, processing, simulation	하강 주제어

<표 5>는 k-평균 군집화를 통해 구분된 8개의 군집을 다시 4가지의 그룹으로 나눈 주제어의 목록을 보여주고 있다. 각각의 그룹은 <그림 1>에서 보여주고 있는 패턴은 '최근 상승 주제어', '상승 & 최근 하강 주제어', '한때 상승 했었던 주제어', '하강 주제어'로 분류하였다.

<그룹 1>에 속한 주제어들은 빈도수가 꾸준히 증가 하다가 최근에는 급격히 증가하는 것을 볼 수 있으며 <그룹 2>에 속해 있는 주제어들은 2000년대 들어 빈도수가 점차 감소하고 있음을 보여주고 있다. 한편, <그룹 3>에 속한 주제어들은 과거 70~80년대에 빈도수가 가장 높았다가 점차 감소하고 있는 추세를 보이고 있으며, 마지막으로 <그룹 4>에 속한 주제어들은 꾸준히 빈도수가 하강 하고 있는 것을 보여주고 있다. 논문의 초록이 연구의 요약정보를 담고 있다는 점을 감안했을 때, <표 5>는 연구자들에게 꾸준한 사랑을 받아온 주제어들, 또는 서서히 관심을 잃어가거나 최근 뜨겁게 탐구되는 연구 분야를 보여주고 있다.

이와 같이 시간에 따른 주제어 군집 결과는 DBLP의 연구동향을 알아볼 수 있는 매우 흥미로운 자료로 사용될 수 있을 것이다.

5. 향후 계획 및 결론

본 연구에서는 컴퓨터공학 분야의 대표적 사이트인 DBLP의 정보를 이용하여 지난 60여 년간 출판된 논문의 연대별 주제어의 출현 패턴을 살펴보았다. k-평균 군집화를 활용하여 유사한 특성을 보이는 주제어들을 군집화 함으로써 흥미로운 자료를 도출했다.

논문제목 및 초록에 포함된 키워드를 분석함으로써, 과거에서 현재에 이르는 DBLP 내 연구의 현황과 추이를 관찰할 수 있었다. 향후에는 논문별 주제어 출현 빈도를 추출해 주제어 간의 연관관계를 알아보고, k-평균 군집분석기법 이외에도 사회관계망 분석 기법 등을 활용해 DBLP의 저자들 간의 관계나 논문의 연관관계를 살펴볼 계획이다.

본 연구는 특정 분야 사이트를 대상으로 분석이 수행되었지만, 본 논문에서 제시한 분석 기법은 다른 분야에도 얼마든지 적용이 가능하다. 예를 들면, 대량의 연예기사에 포함된 연예인들의 상호관계 또는 인터넷 쇼핑몰의 제품 후기에 포함된 주제어 및 주제어 간의 관계를 살펴 볼 수 있을 것으로 판단된다. 본 연구가 웹 마이닝과 텍스트 마이닝 영역의 다양한 연구를 활성화 시킬 수 있기를 기대해 본다.

참고문헌

[1] Feldman, Ronen and James Sanger. "The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data." DAGLIB, 2007.

[2] A. Kao and S. R. Poteet, Grid Natural Language Processing and Text Mining, Springer, London, UK, 2007.

[3] Yuen-Hsien Tsenga, Chi-Jen Linb, and Yu-I Linc, Text mining techniques for patent analysis, Information Processing & Management, Vol. 43, No. 5, pp. 1216-1247, Sept. 2007.

[4] S. Lee, S, Seol, H.: Using patent information for designing new product and technology: keyword based technology roadmapping, R and D Management, pp. 169-188, 2008.

[5] Su-Gon Cho and Seoung-Bum Kim, Identification of Research Patterns and Trends through Text Mining, International Journal of Information and Education Technology, Vol. 2, No. 3 June (2012)

[6] Su-Gon, Cho., Seoung-Bum, Kim, Finding Meaningful Pattern of Key Words in IIE Transactions Using Text Mining, Journal of the Korean Institue of Industrial Engineers (2012)

[7] J. Poelmans, D. I. Ignatov, S. Viaene, G. Dedene, and S. O. Kuznetsov, Text Mining Scientific Papers: A Survey on FCA-Based Information Retrieval Research, Vol 7377, pp. 273-287, Lecture Notes in Computer Science, 2012

[8] P. Jomsri and D. Prangchumpol, A hybrid model ranking search result for research paper searching on social bookmarking, Proceedings of INISCom 2015, pp. 38-43, 2015.

[9] DBLP, http://dblp.uni-trier.de

[10] Gartner IT term, http://blogs.gartner.com/it-glossary