**데이터 마이닝 최종 프로젝트 결과 보고서**

제출일 2021.05.20

제출인: 12조 2019314680 손세정

12조 2019314962 안채희

<차례>

1. 서론
   1. 연구의 필요성 및 목적
   2. 연구 주제 수립
2. 연구 절차 및 분석 방법

2-1. 데이터 수집 단계

2-2. 데이터 전 처리 단계

2-3. 데이터 시각화 단계

1. 연구 결과 및 해석
2. 요약 및 결론

4-1. 연구 결과 요약

4-2. 연구 활용과 향후 연구 방향 제시

1. 참고 문헌
2. **서론**

**1-1. 연구의 필요성 및 목적**

본 연구는 검색 기능에만 치중하여 분야별 변천사와 분야간 관계성을 파악하기 힘든 기존 논문 사이트의 문제점을 해결하고자 하였다. 이러한 문제로 학습자는 어떤 논문을 왜 읽어야 하는지, 그 논문이 어떠한 학문적 맥락에서 중요한지 아무런 배경지식이 없는 체로 읽게 된다. 즉, 세부 분야별 연구 주제의 변천사를 파악하기 위해 세부 분야를 연도별로 검색해야 하는 어려움과, 분야별로 대두되는 주제 또한 따로 조사해야 하는 불편함을 가진다. 연구자 역시 연구 전 사전 조사에 있어서 각 분야별 자주 연구된 주제와 연구가 부족한 주제 중 선정하고자 하는 주제에 대한 판단을 위해 각 분야별 키워드 지표를 개인이 조사해야 한다. 뿐만 아니라, 연구하고자 하는 주제와 연관성이 높은 타 분야를 참고할 때 지나치게 방대한 양의 논문으로 인하여 어려움을 겪는다. 그러므로 이러한 문제점을 해결하기 위해 본 연구는 컴퓨터 공학의 연도별 세부 분야의 연구주제 변천사와 분야간 연관성을 연도 기준으로 통시적으로 확인할 수 있도록 시각화 지표를 형성했다. 나아가 공시적 관점에서 연도별로 어떠한 키워드가 크게 연구되기 시작했는지 시대적, 사회적 맥락에서 분석하여 연구 주제의 트렌드와 사회문화적 영향을 알고자 했다.

**1-2. 연구 주제 수립**

이러한 본 연구의 목적에 맞게 근 21년간 컴퓨터 공학 분야 내의 저명한 논문들을 활용하여 각 논문의 분야와 참고문헌으로 인용된 논문의 분야를 연결하고, 네트워크로 시각화하였다. 이를 통해 각 시기별로 어떠한 분야 사이에 큰 연관성이 있었는지, 그 연관성이 시대에 따라 어떻게 변화하였는지 직관적으로 파악할 수 있도록 하였다. 또한 연도별 각 분야의 인용 점수와 키워드를 통해 전체적으로 어떤 연도의 연구 주제가 해당 주제 및 타 연구에서 자주 사용되었는지 알아보고 시대별 크게 화제가 되었던 사건들과 연관 지어 분석하였다. 따라서 본 연구에서는 컴퓨터 공학 내 방대한 논문 정보와 분야간 연관성을 각 척도에 맞게 시각화하여 제공하고, 이를 통시적 및 공시적 측면에서 분석하고자 하였다.

1. **연구 절차 및 분석 방법**

**2-1. 데이터 수집 단계**



해당 연구에 사용할 데이터는 세계 최초의 컴퓨터 공학 학회 연합체인 ACM의 Digital Library 사이트에서, 컴퓨터 공학 내의 세부 학문 분야들로 구분하여, 위와 같은 13개의 카테고리로 나누어 수집하였다. 논문의 인용 횟수를 해당 논문의 공신력을 수치화하는 기준으로 삼아, 각 카테고리별로 2000년부터 2020년까지 연도별로 인용 횟수 상위 300개 내외를 수집하여 csv 파일의 형태로 저장하였다. 한 논문 당 해당 논문의 종합적인 정보가 들어있는 BibTex 인용 양식과, 해당 논문이 인용하고 있는 ACM 논문들, 해당 논문을 고유하게 식별할 수 있는 DOI 번호를 column 값으로 지정하여 총 61000여개의 논문 데이터를 이 연구에 활용하였다.

**2-2. 데이터 전 처리 단계**

수집했던 데이터들 중에 각 논문의 고유 식별 번호인 DOI 값이 nan인 경우 해당 논문은 삭제하였다. 또 각 데이터의 열 값 중 BibTex 인용 양식에서, 각 논문들을 대표하는 특징으로 선정한 title과 abstract, keyword, author, year 총 다섯 개의 정보를 새로운 열로 추가했다. Keyword의 경우 BibTex에 없다면 gensim의 keyword 추출 기능을 사용하여 abstract에서 대표 keyword를 추출해 내었고, 만약 abstract도 존재하지 않는다면 nltk의 pos tagging과 word tokenize을 활용하여 논문의 title에서 동사, 명사, 형용사 등 유의미한 token들을 keyword로 선정하였다.

**2-3. 데이터 시각화 단계**

연구의 주제와 목적에 부합하는 데이터 시각화 그래프와 네트워크를 크게 컴퓨터 공학 전체 논문 데이터 인용 네트워크, 각 학문 분야별 연도 간 핵심 keyword word Cloud와 선 그래프, 연도별 컴퓨터 공학 카테고리 간 citation score 비교 차트로 세 종류로 나누어 제작하기로 하였다.

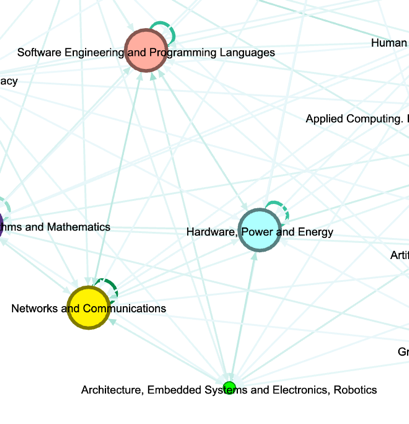
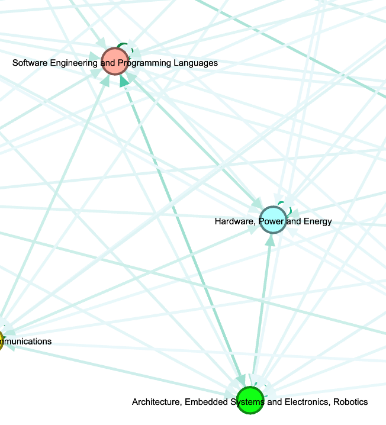
첫번째로, 컴퓨터 공학 전체 논문 데이터 인용 네트워크는 수집한 모든 데이터를 바탕으로 논문 인용 관계를 파악하여, 컴퓨터 공학 내의 세부 카테고리 간에 어떠한 상호작용과 긴밀성을 가지고 있는지 가시적으로 확인하기 위하여 제작하였다. 네트워크 시각화 도구로서는 Gephi 프로그램을 사용하여 다양하게 정보를 담아내고자 했다. 노드는 ACM에서 제공한 카테고리로 구별되며, 노드의 크기는 참고 문헌으로 사용된 횟수와 비례하도록 하였다. 네트워크의 엣지는 인용한 카테고리에서 인용된 카테고리로 향하도록 단방향으로 연결하였고, 각 연도에 해당 카테고리가 다른 각 카테고리의 논문에서 인용된 총 횟수를 기준으로 비례하게 선명도를 조절하였다.

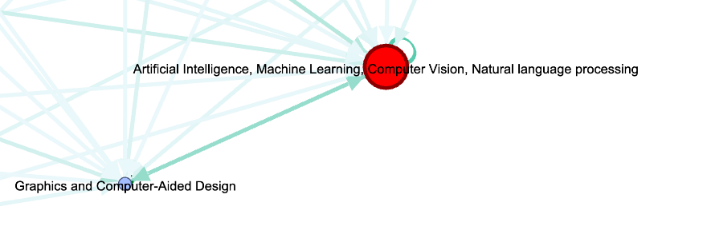
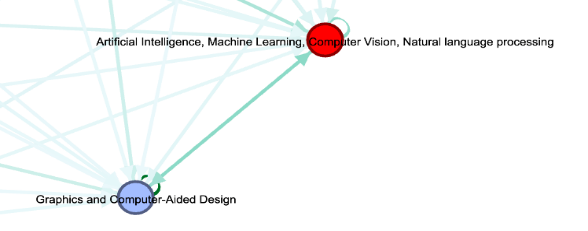
두번째로, 통시적인 관점에서 컴퓨터 공학 내의 각 세부 분야별 연도 간에 가장 많이 등장하는 핵심 keyword 상위 10개를 통해, 어떤 주제가 해당 카테고리 내에서 가장 이슈가 되었는지 그 변화를 확인해보고자 한다. 이는 해당 학문 내의 획기적인 아이디어의 등장이나 연구 트렌드의 흐름을 유기적으로 파악하기에 유리할 것이라 판단하여 word Cloud를 활용하고자 했다. 또한 가시적인 확인을 위하여, 단어의 등장 횟수가 연도에 따라 어떻게 변화하는지 확인하기 위해 선 그래프 또한 제작하였다. 가공한 데이터에서 keyword, category, year 열을 활용하여, 모든 연도에 공통으로 상위 5개 내에 든 단어들은 해당 분야를 전반적으로 잘 나타내는 단어로 판단하였다. 또한 그 keyword들은 제거하여 학문 내의 트렌드 변화를 뚜렷하게 관찰하고자 하였다

세번째로, 공시적 관점에서는 최근 21년 동안의 13개의 카테고리 간 citation score를 비교하는 차트를 통하여, 같은 시기의 컴퓨터 공학 내에서 어떤 세부 학문의 영향력이 더 확대됐고 축소되었는지 확인하고자 시각화했다. Citation score는 해당 논문의 인용 횟수로, 그 논문이 공신력 높고 신뢰할 수 있음을 증명하는 척도로 판단했다. 특히 citation score 비교 차트들을 통해 어떤 세부 분야가 컴퓨터 공학 내에서 가장 큰 연구 주제가 되는지, 어떤 세부 분야가 부상하는지, 또는 그것이 시대적인 상황에 영향을 받았는지를 확인할 수 있는 좋은 시각화 자료가 될 것이라 판단했다. 또한 시각화 도구로 Tableau를 사용하였으며 차트의 x축은 연도와 카테고리로, y축은 각 카테고리의 citation score는 해당 연도의 해당 카테고리의 논문들의 인용 횟수의 합으로 설정했다.

1. **연구 결과 및 해석**

먼저, 각 카테고리를 인용 여부를 기준으로 타 카테고리와 연결시키고, 연결된 에지의 개수 및 총 인용된 횟수로 네트워크를 조정하여 시각화하였을 때, 연도별로 각 카테고리 사이의 연관성을 한눈에 알아볼 수 있게 되었다. 이를 활용하여 분석한 결과 각 카테고리의 연결도가 해마다 변화하기도, 일정하게 밀접한 관계를 맺기도 하는 것을 알 수 있었다. 이는 Architecture, Embedded Systems and Electronics, Robotic (이하 Architecture)와 Software and Programming Languages (이하 Software) 사이의 연결도, Artificial Intelligence, Machine Learning, Computer Vision, Natural language processing (이하 Artificial Intelligence)와 Graphics and Computer-Aided Design (이하 Graphics) 사이의 연결도를 통해 명확하게 확인할 수 있다.

 <2003년도 네트워크 일부> <2012년도 네트워크 일부>

전자의 경우 Architecture와 Software 카테고리는 2003년부터 2007년까지 에지의 선명도가 약해 각 카테고리 간에 인용 횟수가 매우 적었음을 알 수 있다. 즉, 각 카테고리 간의 연결도가 타 연도에 비해 적었다고 할 수 있다. 반면에 2008년부터 2015년의 네트워크를 분석하였을 때 Architecture가 Software 내 논문을 인용하는 횟수가 눈에 띄게 증가하였음을 알 수 있다. 그러나 그 후부터 2020년 사이에 점점 에지가 희미해져 점차 연결도가 약해지는 것을 통해 시대에 따라 세부 분야가 각 연구에서 그 연결 정도가 유동적으로 변화함을 알 수 있다.

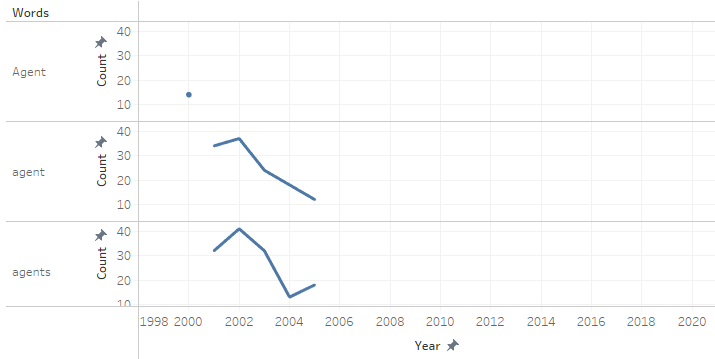
<2008년도 네트워크 일부> <2012년도 네트워크 일부>

후자의 예시는 전체적인 시대 흐름과 별개로 항상 밀접한 연관성을 보인 경우로, 위 네크워크의 일부는 Artificial Intelligence와 Graphics 사이의 연결도가 양방향으로 선명하게 연결되어 있음을 보여준다. 두 카테고리는 2000년도를 제외한 모든 연도에서 비슷한 수준의 선명도를 가진 양방향 에지로 연결되어있어 한 분야가 다른 분야를 인용하는 것이 아닌, 양 분야 모두 서로의 연구를 인용하는 것으로 해석할 수 있다. 즉, 두 세부 분야의 전반적인 연구가 긴밀하게 연결되어 있는 것으로, 이는 특히 Artificial Intelligence 내 Natural language processing을 제외한 모든 주제가 Graphics 카테고리 내 주제와 높은 연관성을 보이기 때문으로 분석할 수 있다.

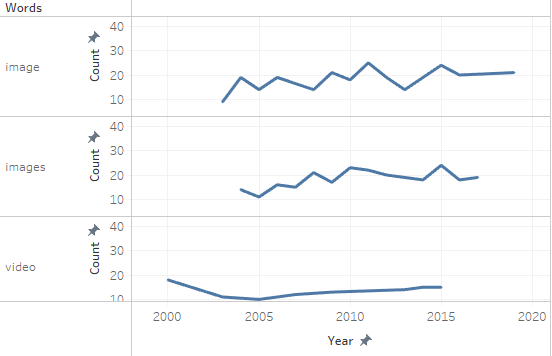
그러므로 연구자는 연구 전 사전 조사에 있어서 각 연도별 네트워크를 참고해 연구하고자 하는 분야가 어떤 연도에, 어떤 학문과 높은 연관성을 보이는지 분석해보아야 한다. 전자의 예시와 같은 연결도를 보이는 분야 간에는 아래의 word Cloud 및 키워드 그래프를 참고하여 해당 연도의 주제 트렌드에 대해 파악할 필요성이 있고, 후자의 예시와 같은 경우 두 분야를 반드시 함께 연구해야 하므로 사전 조사 및 계획에 있어서 두 분야를 모두 포함해야 한다.

다음으로, 각 학문 분야별 연도 간 핵심 keyword word Cloud와 선 그래프 시각화 결과, 해당 분야 내에서의 트렌드 변화를 통시적으로 관찰할 수 있었다. 한 예시로, 실제 연구 결과 중 Artificial Intelligence, Machine Learning, Computer Vision, Natural language processing 분야에 대한 학문 키워드의 변화 과정을 다음과 같이 확인할 수 있었다.

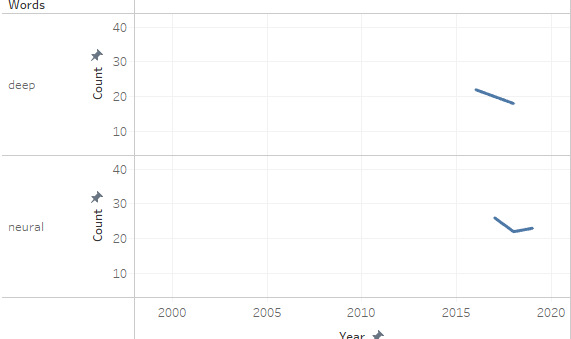


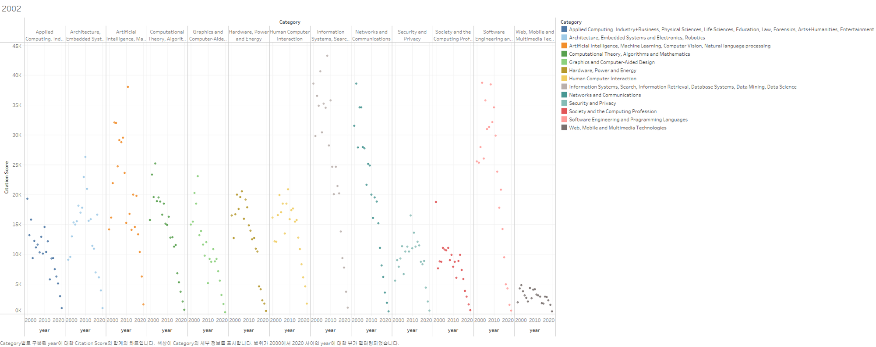
실제로 Artificial Intelligence, Machine Learning, Computer Vision, Natural language processing 분야에서 2000년부터 2005년까지 상위에 차지하고 있는 keyword는 agent로, 인공지능 분야에서의 Agent란 복잡한 실제 환경 속에서 특정 목적을 달성하기 위해 사용자를 대신하여 작업을 수행하는 시스템을 지칭하는 인공지능의 한 연구 분야다. agent라는 keyword는 2005년 이후 더 이상 활발히 연구되지 않았음을 알 수 있다.



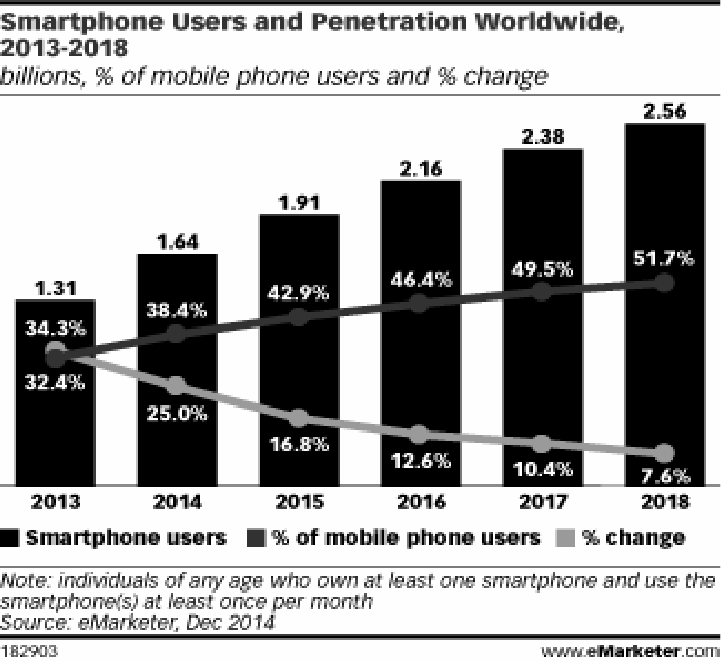
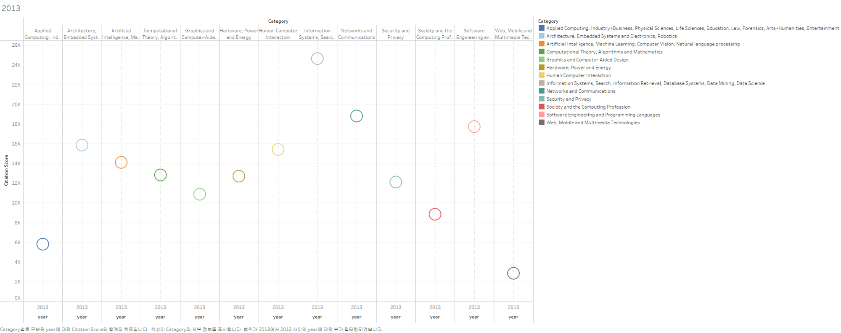
그에 반해 2004년부터 차츰 비중이 커진 image라는 keyword는 2003년부터 확산되어 2006년부터 큰 비중을 차지하다가 2009년부터 2014년까지 인공지능 관련 분야에서 가장 활발히 연구되는 주제였음을 확인하였다. 이는 2001년 9.11 테러 이후 국가 보안을 위해 범죄자 얼굴 인식의 필요성이 높아지면서 face recognition을 중심으로 image 분석이 발전하기 시작하였다. (Kelly Gates 2006) 뒤이어, 스마트폰의 공급과 함께 고화질의 이미지를 쉽게 대량으로 얻을 수 있게 되었으며, (OZCAN 2017) 2010년 당시 ImageNet Large Scale Visual Recognition Challenge이 시작됨에 따라 이미지 관련 컴퓨터 비전이 인공지능이란 학문 내에서 가장 이슈가 되었다는 것을 연관 지어 해석할 수 있었다.



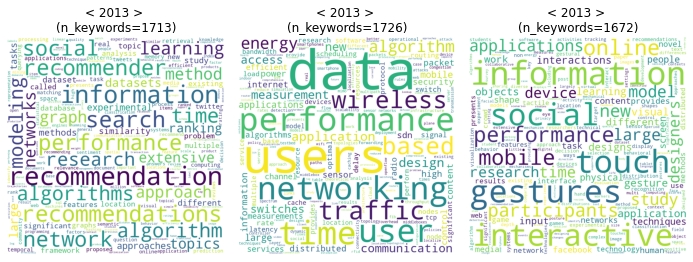
이 영향으로 딥 러닝이 2012~2013년에 컴퓨터 비전과 지각에 관련된 모든 문제에서 최고의 성능을 내자 2014년 사이에 2016년부터 batch normalization, residual connection 등 성능을 향상시키는 다양한 방법들이 고안되었으며 deep learning과 neural network는 2019년까지도 학문 내에서 가장 높은 관심사가 되었음을 word cloud와 선 그래프를 통해 통시적으로 파악할 수 있었다.



마지막으로, 시간의 흐름에 따라 13개의 카테고리 간 citation score를 비교하는 차트 결과를 통해, 컴퓨터 공학 내의 어떤 세부 학문의 영향력이 더 확대되었고 축소되었는지 확인할 수 있었다. 최근에 발행된 논문의 citation score는 상대적으로 낮을 것이라는 가정 하에 연도별로 비교하여 해석하였다. 먼저 모든 해에 거쳐서 컴퓨터 공학 내의 연구 분야 중, Applied Computing: Industry/Business, Physical Sciences, Life Sciences, Education, Law, Forensics, Arts/Humanities, Entertainment과 Software Engineering and Programming Languages와 Networks and Communications, 그리고 Information Systems, Search, Information Retrieval, Database Systems, Data Mining, Data Science 이렇게 네 분야가 근 21년 간 컴퓨터 공학 분야에서 가장 영향을 많이 미쳤다고 해석할 수 있다.



그중 2013년의 citation score 비교 차트를 위의 keyword cloud와 함께 분석하고자 한다. 2013년에 citation score가 가장 높았던 컴퓨터 공학 세부 분야는 Information Systems, Search, Information Retrieval, Database Systems, Data Mining, Data Science 로, 이때 정보와 데이터 분야에서 가장 이슈가 되었던 연구 keyword는 recommendation, recommender, recommendations 등 추천과 관련된 내용이었다. 그 다음 높은 citation score를 가진 Networks and Communications 분야에서는 networking과 wireless와 같인 무선 네트워크와 관련된 keyword가 다수 등장하였다. 2013년 Human Computer Interaction 분야에서는 social, touch, gestures, mobile 와 함께 사람의 손동작과 관련된 키워드가 가장 많이 등장하였다. 이를 통해, 세계적으로 스마트폰으로의 교체 비율이 34.3%로 가장 높은 시기, 즉 보급 속도가 가장 빠른 시기에 맞게 2013년의 컴퓨터 공학 내의 학문 영향력과 그 키워드가 유기적인 상관 관계가 있음을 공시적으로 확인해 볼 수 있었다.



1. **요약 및 결론**

**4-1. 요약**

컴퓨터 공학 내의 모든 세부 분야들에 대해 학습하거나 연구하고자 할 때, 검색 기능에 치중되어 있는 기존 사이트의 한계를 넘어, 컴퓨터 공학 내 방대한 논문 정보와 분야간 연관성을 시각화하여 제공하고, 각 분야들을 통시적인 관점과 공시적인 관점에서 분석해보았다. 그 결과 컴퓨터 공학이라는 학문 내 논문들이 시대적, 그리고 사회적 맥락을 반영했다는 것을 알 수 있었다.

**4-2. 연구 활용과 향후 연구 방향 제시**

대학생이나 연구자들에게 학문의 발전 과정 및 현재 상황과 연구 방향에 대한 정보를 제공하고 관련 학문이 어떤 다른 학문과 연관이 되었는지를 파악할 수 있도록 하여 연구 주제를 선정하고 방향성을 잡을 때 도움을 줄 수 있을 것이다. 또한 그들에게 시대적, 그리고 사회적 맥락과 연관 짓도록 처음 보는 논문을 공부하거나 연구할 때 배경 지식을 제공하는 것으로써 활용할 수 있을 것이다.

향후에 추가적으로 시 계열을 이용한 데이터 분석을 통해 과거의 논문 주제들 내에서 트렌드의 주기를 파악하여, 앞으로 어떤 분야에 대한 연구 방향이 유망할지 예측하는 예측 모델을 만들어 본다면, 이 연구 목적의 연장선에서 컴퓨터 공학 내 각 분야의 잠재적인 가능성과 새로운 발전을 도모할 수 있을 것이라 기대한다.

1. **참고문헌**

Smartphone Users and Penetration Worldwide, 2013-2018 Source: eMarketer, Dec 2014

Kelly Gates (2006) IDENTIFYING THE 9/11 ‘FACES OF TERROR’, Cultural Studies, 20:4-5, 417-440

OZCAN, KORAY, "Computer Vision Algorithms for Mobile Camera Applications" (2017). Dissertations - ALL. 704