자동 분류 기법과 지적 구조 분석 기법을 융합한 처방적 분석 시스템 구현 방안 연구

Prescriptive Analytics System Design Fusing Automatic Classification Method and Intellectual Structure Analysis Method

정도헌 (Do-Heon Jeong)*

초 록

본 연구는 새로운 분석법으로 떠오르는 처방적 분석 기법을 소개하고, 이를 분류 기반의 시스템에 효율적으로 적용하는 방안을 제시하는 것을 목적으로 한다. 처방적 분석 기법은 분석의 결과를 제시함과 동시에 최적화된 결과가 나오기까지의 과정 및 다른 선택지까지 제공한다. 새로운 개념의 분석 기법을 도입함으로써 문헌 분류를 기반으로 하는 응용 시스템을 더욱 쉽게 최적화하고 효율적으로 운영하는 방안을 제시하였다. 최적화의 과정을 시뮬레이션하기 위해, 대용량의 학술문헌을 수집하고 기준 분류 체계에 따라 자동 분류를 실시하였다. 처방적 분석 개념을 적용하는 과정에서 대용량의 문헌 분류를 위한 동적 자동 분류 기법과 학문 분야의 지적 구조 분석 기법을 동시에 활용하였다. 실험의 결과로 효과적으로 서비스 분류 체계를 수정하고 재적용할 수 있는 몇 가지 최적화 시나리오를 효율적으로 도출할 수 있음을 보여 주었다.

ABSTRACT

This study aims to introduce an emerging prescriptive analytics method and suggest its efficient application to a category-based service system. Prescriptive analytics method provides the whole process of analysis and available alternatives as well as the results of analysis. To simulate the process of optimization, large scale journal articles have been collected and categorized by classification scheme. In the process of applying the concept of prescriptive analytics to a real system, we have fused a dynamic automatic-categorization method for large scale documents and intellectual structure analysis method for scholarly subject fields. The test result shows that some optimized scenarios can be generated efficiently and utilized effectively for reorganizing the classification-based service system.

키워드: 처방적 분석, 지적 구조, 자동 분류, 분류 체계, 최적화
prescriptive analytics, intellectual structure, automatic classification, classification scheme, optimization

* 덕성여자대학교 문헌정보학과 조교수(doheonjeong@duksung.ac.kr)

■ 논문접수일자 : 2017년 11월 14일 ■ 최초심사일자 : 2017년 11월 28일 ■ 게재확정일자 : 2017년 11월 29일
■ 정보관리학회지, 34(4), 33-57, 2017. 〔http://dx.doi.org/10.3743/KOSIM.2017.34.4.033〕

1. 서 론

빅 데이터 개념의 세 가지 핵심 요소인 3V는 대용량(volume), 속도(velocity), 다양성(variety)이다. 다양한 유형(variety)으로 생산되는 대량의 정보 자원(volume)을 신속하게 처리(velocity)하고 효과적으로 관리하기 위해 많은 분석 기반 정보 서비스들이 분류 체계를 활용하고 있다. 기하급수적으로 증가하는 데이터 프로세싱 환경에서 데이터를 처리, 분석한 후 개선점을 도출하고 서비스를 재편하는 일련의 순환 과정을 전문가에만 의존하는 것이 불가능하게 되었다. 또한 분류 성능이 보장되지 않은 상황에서 기계적인 처리 방식의 도입만으로 분석적 서비스를 개선하는 것 또한 한계가 존재한다. 그러나 성능을 보장한다는 조건 하에서 고비용 수작업 방식의 가장 좋은 대안은 자동화된 처리 방식일 것이다.

미 분류된 학술 정보(unlabeled documents)가 대량으로 유입되는 경우에 일관성 있는 방식으로 모든 자원에 분류코드를 부여하는 것이 우선적으로 필요하며, 이를 위해 다양한 자동 분류 기법을 활용할 수 있다(이원구, 신성호, 김광영, 정도헌, 윤화묵, 성원경, 이민호, 2011). 그러나 학술 문헌의 자동 분류의 결과는 인문학, 공학과 같은 학문 계열의 특성, 세부 학문 분야에서 사용하는 용어의 전문성 등의 다양한 요인으로 성능의 편차가 크게 나타나기도 한다(정도헌, 김환민, 김혜선, 신기정, 2007). 텍스트 마이닝 기술을 기반으로 하는 지능형 시스템을 운영하는 경우에는 첫 번째로 분류 정보(labeled data)의 정확성과 신뢰성을 높이는 기술을 확보하고, 두 번째로 전체적인 시스템의 분석 수준을 상시 모니터링하면서 시스템의 프로세싱 체계를 정비하는 피드백 환경이 구비되어야 한다.

첫 번째 방안에 대한 해답은 능동적 학습(active learning)이라는 기계 학습의 새로운 접근법에서 찾을 수 있다. 모든 데이터를 수작업 평가하여 기계에 학습시키는 방식을 수동적 학습(passive learning)이라 일컫는 데 반해, 범주 정보가 없는 데이터가 주어졌을 때 기계가 이를 살펴보고 애매한 것 혹은 경계지점에 놓인 것들을 찾아내 사람이 판단하도록 함으로써 값비싼 범주 데이터를 효과적으로 구축하는 방식이 능동적 학습기법의 특징이다(McCallum & Nigam, 1998; Settles, 2010).

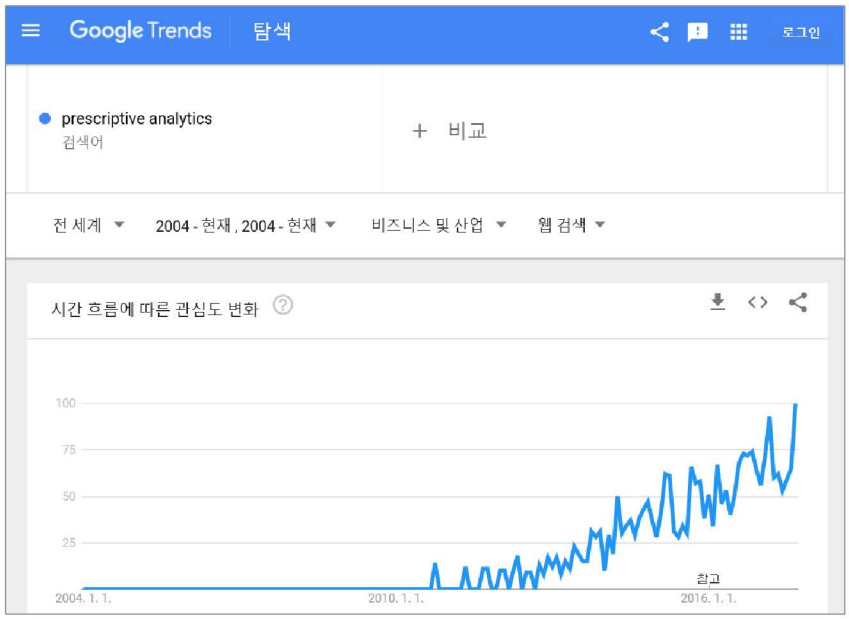
두 번째 방안인 전체적인 시스템의 성능을 분석하여 최적의 방안을 도출하는 방법과 관련하여, 최근 새로운 분석 기법으로 떠오르는 처방적 분석 방법론(prescriptive analytics: PA)을 고려해 볼 필요가 있다. 능동적 학습이 양질의 학습 데이터를 확보함으로써 기계적인 학습 모델의 성능을 개선하고자 하는 것을 목적으로 한다면, 처방적 분석 개념은 주어진 시스템을 최적화하여 성능을 극대화하는 처리 프로세싱 과정을 다룬다. 이 기법에서 중요한 점은 판단의 근거가 되는 이유를 제시한다는 점이다(Bertsimas & Kallus, 2015; Song, Kim, Hwang, Kim, Jeong, Lee, & Jung, 2013). 시스템이 최적화하여 제공하는 처방(prescription)에 대한 이유를 사용자 또는 분석가가 이해할 수 있어야 한다. 이로부터 시스템의 사용자는 합리적인 최종 판단을 내릴 수 있을 것이며, 그 결정에 대한 당위성을 확보할 수 있게 된다. 또한 인공지능 기법을 이용하는 추천 시스템의 결과가 만족스럽지 않을 때, 처방적 분석 개념이 도입된 시스템은 그렇

지 않은 경우보다 훨씬 효과적으로 시스템의 데이터 프로세싱 전 과정을 점검할 수 있다.

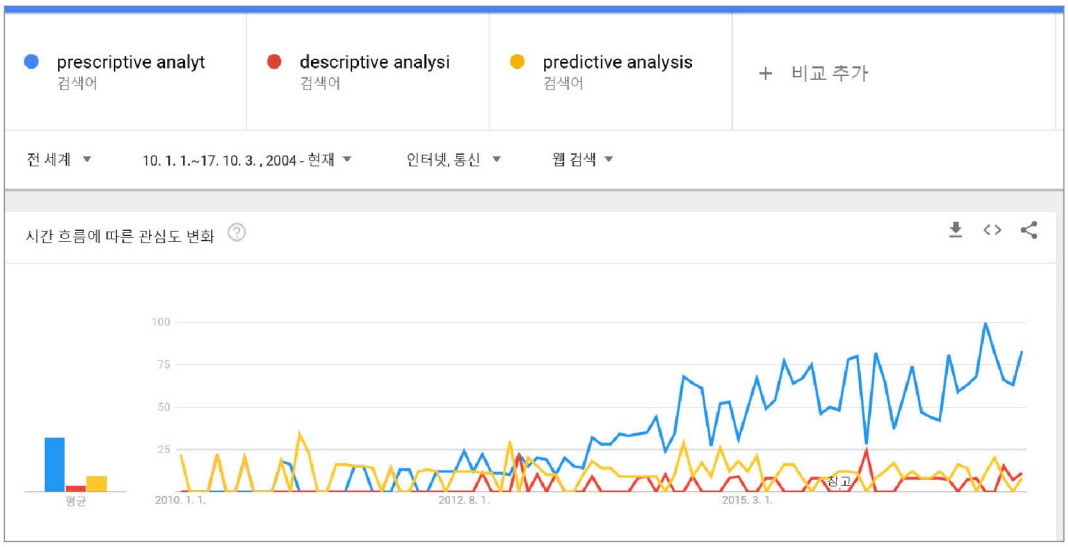
특히 이러한 처방적 분석 방법은 보다 나은 선택을 위한 옵션을 제공한다는 장점으로 인해 비즈니스 및 산업 분야에서 활발히 연구가 이루어지고 있으며, 〈그림 1〉에서 보는 바와 같이 2010년 이후 인터넷 상의 정보가 급격하게 증가하고 있는 추세이다. 또한 전통적인 분석 기법으로 널리 사용되어 왔던 기술적 분석 기법(descriptive analysis)이나 현재 많이 사용되는 예측 분석 기법(predictive analysis)과 비교해 보아도 처방적 분석 기법(prescriptive analytics)의 최근 양적 강세가 두드러짐을 확인할 수 있다(〈그림 2〉 참조).

본 연구에서는 다음의 몇 가지 이슈들을 연구의 배경으로 다루고 있다. 대용량 데이터에 기반한 지능형 서비스를 위해서는 고성능의 분류 기술 도입이 요구되고 있다. 더욱이 학제 간 연구가 활발해지고 문헌이 많이 구축되면 자동 분류 성능이 현재보다 더욱 저하될 수 있다. 실제 현업에서 분류 기반 서비스의 이용자 만족도가 낮은 경우, 기존 분류체계를 조정할 당위적 근거 마련이 어려운 점 또한 해결해야 할 과제이다. 데이터를 효과적으로 재구성하기 위해서는 전체 데이터의 분야별 분포를 시각적으로 해석할 수 있는 방법론이 필요하며, 서비스를 개선하기 위해 자동 또는 반자동으로 시스템이 제공하는 진단 정보가 필요할 것으로 보인다.

본 연구는 이러한 연구 주제들을 배경으로 처방적 분석 기법에 기반한 지능형 서비스 설계를 목표로 한다. 처방적 분석 개념을 적용하



〈그림 1〉 2010년 이후 급격히 증가하는 처방적 분석 기법
(구글 트렌드, 2017년10월4일자)



〈그림 2〉 기술적 분석, 예측 분석, 처방적 분석 기법의 전세계 웹 정보량
(구글트렌드(인터넷 & 통신분야), 2017년10월4일자)

기 위해, 우선 대용량의 문헌을 빠르게 학습하는 데 특화된 동적 분류기술을 도입하고자 하였다. 또한 구축된 학술정보의 주제 범주 간 특성을 파악하기 위하여 네트워크 척도 알고리즘(network scaling algorithm)을 이용해 학문의 전역적 지식 구조를 분석하고자 하였다.

다음 2장에서는 처방적 분석 시스템 설계를 위한 관련 연구와 선행 연구를 제시한다. 대량의 문헌을 작은 학습 셋으로 분할하여 빠르게 학습하면서도 자유롭게 학습 셋의 규모를 조작할 수 있는 동적 학습 방식의 분류기법을 설명한다. 또한, 전역적인 학문 구조를 파악하기 위한 기법으로 계량정보학(informetrics)의 개념과 네트워크 척도 알고리즘인 패스파인더 네트워크(pathfinder network) 알고리즘을 소개한다. 마지막으로 발전된 분석 기법인 처방적 분석에 대한 개념을 설명하고 관련 시스템 개발의 사례를 언급한다. 3장에서는 처방적 분석 시스템이 도출하는 시나리오를 작성하기 위한 일련의 실험 과정을 설명한다. 우선, 데이터의 수집과 분류 체계 설정을 다루고, 학술논문의 자동 분류의 과정을 통해 학문 분야 간 연관정보를 얻는 방법을 설명한다. 자동 분류 과정에서 생성된 오류로그 파일을 활용하여 학문 분야의 지적 구조를 도출한다. 마지막으로 정보 시각화를 통해 주요 학문영역 간의 지적 구조를 해석해 본다. 4장에서는 처방적 시스템을 위한 시나리오를 도출하는 과정을 설명한다. 지적 구조를 변경함으로써 시스템의 분류 서비스를 개선할 수 있는 몇 가지 처방적 시나리오를 도출하여, 시각적 분석 자료가 주제 분야별 성능 편차의 원인을 파악하고 주제 범주를 재설정하는 일련의 프로세스에 있어 합리적 의사 결정에 효과적임을 살펴보고자 한다. 5장에서는 연구의 의의와 한계점을 언급하고, 향후 연구에 대해 간단히 소개한다.

2. 선행 연구 및 관련 연구

2.1 처방적 분석에 적합한 자동 분류 기법

처방적 분석에 기반한 분류 시스템을 구축하기 위해서는 처방적 분석 개념을 지원할 수 있는, 적절한 기계학습 기법을 선정하는 것이 중요하다. 처방적 분석은 여러 가지 선택지를 제공하는 다양한 시나리오 생성을 해야 하기 때문에 대량의 데이터에 대해 학습 셋을 다양하게 변화시키면서 반복적인 학습 작업을 빠르게 수행할 수 있는 기법이 필요하다. 즉, 대용량 문헌의 학습에 제약이 없어야 할 뿐 아니라, 최적의 옵션을 찾기 위해 반복적으로 데이터 셋을 나누고 결합하는 방식에 적합한 분류기를 구비하는 것이 필수적이다.

본 연구에서는 대용량 문서에 적합한 동적 문헌 분류 기법으로 자질값 투표형 분류기(Feature Value Voting Classifier: FVC)를 사용한다. FVC는 데이터 집합을 최소 단위로 나누어 학습하고 필요시 작은 단위로 학습된 결과를 즉시 재결합하는 방식으로 대량의 데이터를 빠르게 학습한다. 문서로부터 추출한 개별 자질에 대한 주제 범주와의 유사도 벡터정보를 생성한 후 문서의 색인어에 다수결 방식으로 투표하여 최다 득표한 주제범주를 선택하는 투표방식 분류기이다(이재윤, 2005; 정도헌, 2010; Jeong, Kim, Hwang, Song, & Jung, 2012; Ko & Seo, 2004). 자질의 주제-가중치 벡터 생성을 위한 유사계수는 다양하게 사용할 수 있으며, *Cosine* 유사계수와 *logTF*IDF* 자질 가중치 요소를 포함하여 공식(1)과 같이 표현할 수 있다. 이때 문헌 벡터 $\vec{d}$는 n개의 자질에 대해 자질 f_i가 범주 c_j에 대해서 가지는 자질 값인 $vs(f_i, c_j)$로 구성되며 공식(2)와 같이 표현할 수 있다. 마지막으로 자질값 투표형 분류기는 다수결 투표를 통해 공식(3)을 만족하는 최다득표 범주 c_j를 최종 분류 결과로 할당한다. 구조적으로 매우 간결하므로 메모리를 적게 차지하고 고속으로 문헌을 분류하는 장점을 지닌 분류기법이다.

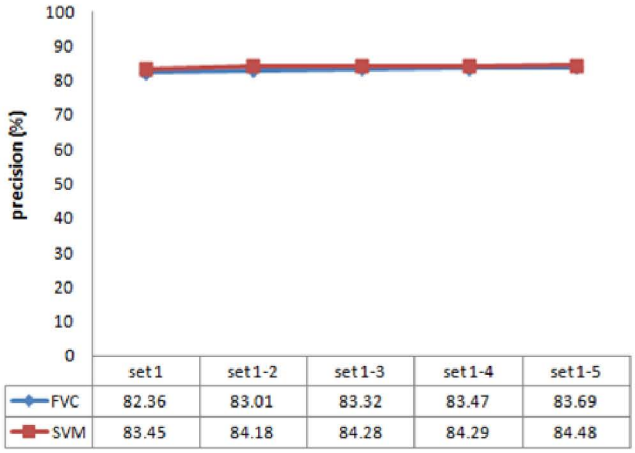
$$vs(f,c_j) = (1+\log tf)\times \log(N/df)\times Cos(f,c_j) \quad (1)$$

$$\vec{d} = \{vs(f_1,c_j), vs(f_2,c_j), \ldots, vs(f_n,c_j)\} \quad (2)$$

$$decision[c_j] = \arg max_{c_j \in C} \sum_i vs(f_i, c_j) \quad (3)$$

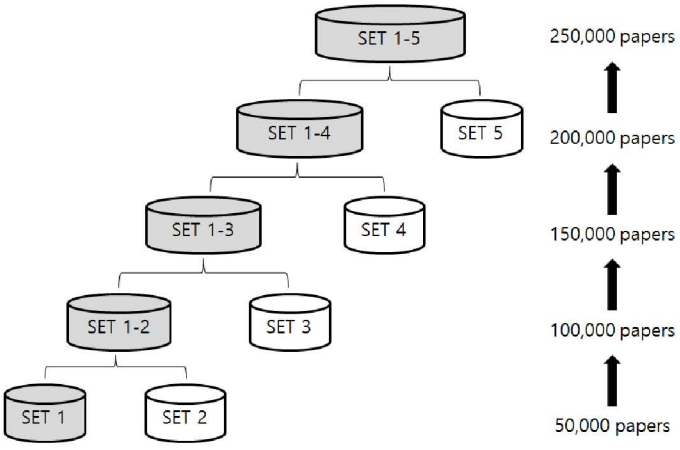
이러한 자질 투표방식 분류기들은 비교적 간단한 구조에 비해 매우 우수한 성능을 보여준다. Ko와 Seo(2004)는 본 연구에서 사용하는 FVC 방식과 유사한 자질값 투영기법(feature projection technique)을 이용한 투표형 분류기인 TCFP(Text Categorization using Feature Projections)를 제안하고 k-NN, Rocchio, Naive Bayes 분류기 등과 비교하여 속도와 성능면에서 우수한 결과를 나타낸 바 있다.

Jeong et al.(2012)은 분할된 학습 데이터 셋들의 동적 결합을 통해 대용량 문헌을 고속으로 분류하는 동적 분류기를 제안하였다. FVC 기반의 동적 분류기와 SVM(Support Vector Machines)과의 성능 비교를 통해 학습 방식과 속도 면에서 장점을 가지면서도 성능 면에서 대등한 결과를 보여주었다(<그림 3> 참조). 작은 단위의 학습용 문헌 집단을 각각 학습한 후 마치 블록으로 쌓듯이 대용량의 학습 결과를 빠르게 재 생성하는 증분 학습(incremental learning)

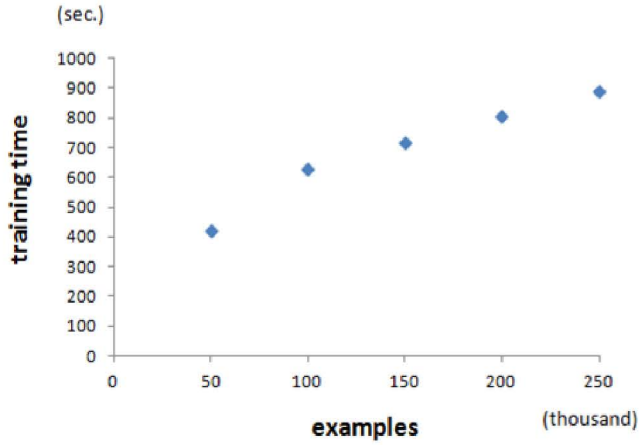


〈그림 3〉 FVC와 SVM 알고리즘 기반 분류기의 마이크로 정확률(micro averaged precision) 비교(Jeong et al., 2012)

알고리즘은 최적의 옵션을 찾아가는 과정에서 수많은 시뮬레이션을 시도해야 하는 처방적 분석 시스템의 필수적 기반 모델이 된다. 〈그림 4〉는 대용량의 문헌을 고속으로 학습하기 위해 계층적으로 쌓아하는 증분 학습의 방식을 설명한다. 이러한 결합 방식을 이용하면 학습 셋을 순차적으로 쌓거나 또는 원하는 학습 셋만 선별적으로 결합하는 등 학습 결과를 원하는 방식으로 무제한 생성할 수 있다는 장점이 있다. 또한 학습 데이터가 증가해도 학습 소요 시간이 일정하게 선형 증가하는 장점이 있기 때문에 대용량 문헌의 학습에 매우 효과적인 모델이라 할 수 있다(〈그림 5〉 참조).



〈그림 4〉 학습 데이터 셋의 계층적 구축 방법(Jeong et al., 2012)



〈그림 5〉 문헌 규모에 따른 FVC 알고리즘의 학습 소요 시간(Jeong et al., 2012)
(Intel Core i7 930@2.8GHz, 8GB Ram)

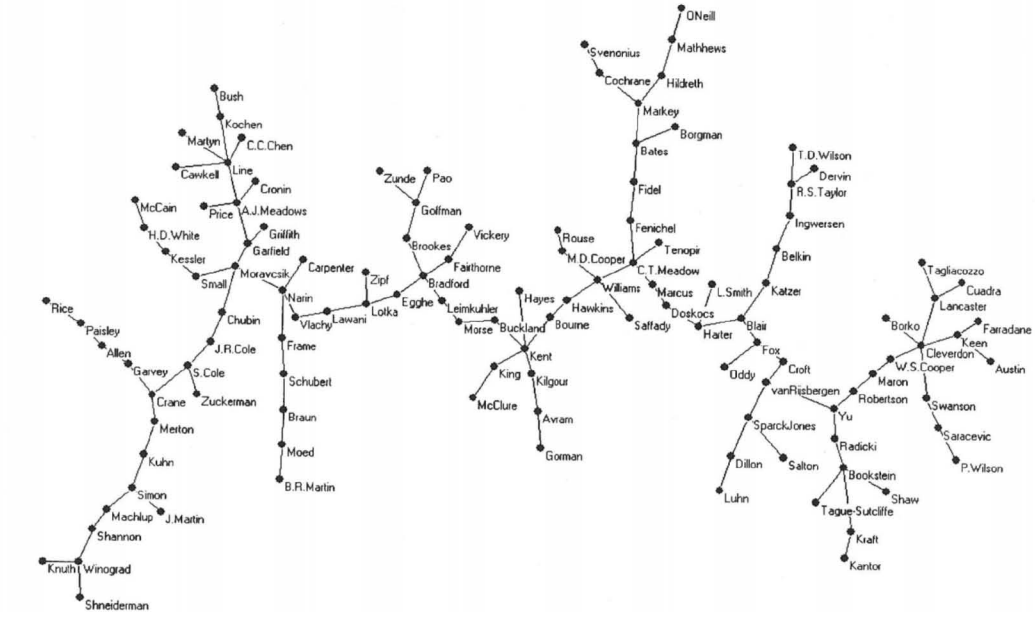
2.2 학문 분야의 지적 구조 분석을 위한 패스파인더 네트워크 척도

지능형 PA 시스템을 구성하는 중요한 기술 요소 중 하나는 시각화이다. 시각적으로 제공되는 각종 자료는 분석의 결과가 도출되기까지의 시스템 처리과정을 사용자가 인지적으로 확인하고 판단의 근거로 삼기 위한 유용한 도구가 될 수 있다. 특히 서로 긴밀하게 연결되어 있는 학문 분야의 분류 정보를 해석하기 위해서는 상호 관계를 보다 구조적으로 시각화할 수 있는 방안인 전역 데이터의 지적 구조 생성 방법이 효과적일 수 있다. 이러한 학문 영역의 지적 구조를 파악하고자 하는 분석적 연구는 오래전부터 계량정보학(informetrics) 분야에서 발전되어 왔다. 계량정보학은 각종 기록물과 문헌정보, 연구자 또는 사회 그룹(social group) 네트워크 등 다양한 유형의 정보를 정량적인 관점에서 분석하고 해석하는 학문으로써(Tague-Sutcliffe, 1992), 계량정보 분석을 통해 연구자 중심의 공동 연구와 인용 관계를 파악하거나(Chua & Yang, 2008; Hou, Kretschmer, & Liu, 2008; White, 2003), 국가별 또는 산·학·연(triple helix) 간의 주요 연구 동향과 추이를 비교 분석하기도 한다(Glänzel & Schilemmer, 2007). 특히, 다학문간(multidisciplinary) 또는 학제간(interdisciplinary) 연구의 수준을 분석하는 연구는, 학문 영역의 전역적인 지적 구조를 파악하기 위해 계량정보학 기법을 이용하여 네트워크를 시각화하고 분석하려는 지식 영역 분석(knowledge domain analysis) 연구로써 계량정보학의 주된 연구 중 하나이다(Kim & Lee, 2008; Levitt & Thelwall, 2008; Morillo, Bordons, & Gomez, 2003). 이밖에도 계량정보학적 분석은 특정연구 분야에서 떠오르는 신규 기술을 찾아내거나 거시적 관점에서 국가의 유망 기술을 예측하는 등의 미래 연구에도 많이 활용되고 있다(Åström, 2007; Zhao & Strotmann, 2007).

본 연구에서는 처방적 분석 기법을 지원하

기 위한 시각화 기법으로 패스파인더 네트워크(PathFinder Network: 이하 PFNet) 척도 알고리즘을 제안한다(Quirin, Cordon, Guerrero-Bote, Vargas-Quesada, & Moya-Anegon, 2008; Schvaneveldt, 1990). PFNet 기법은, 기존의 다차원 척도법(Multidimensional Scaling: MDS)이 지적 구조를 세부적으로 표현하는 데 어려움이 있어 이를 보완하는 네트워크 기반의 알고리즘이 주목받으면서 최근 계량학적 분석 연구에서 활발하게 사용되는 네트워크 척도 알고리즘 중 하나이다(〈그림 6〉 참조).

〈그림 7〉은 패스파인더 네트워크 척도의 구현 알고리즘을 설명한다. PFNet은 두 개의 파라미터를 사용하는데, q는 노드 간 경로거리를 측정하기 위한 최대 링크의 수를 뜻하며, r은 민코프스키 거리(Minkowski metric) 공식의 제곱수로서 경로 구성 링크의 가중치를 거리에 반영하는 방법을 뜻한다. PFNet을 구현하기 위해서는 두 개의 매트릭스(W^i_{jk} 와 D^i_{jk})가 필요하다. W^i_{jk} 는 i개의 링크를 따라 노드 j가 노드 k로



〈그림 6〉 PFNet 알고리즘으로 그린 동시인용 기반 저자 네트워크 구조(White, 2003)

1. Compute $W^{i+1} = W \Theta W^i$, as follows: $w^{i+1}_{jk} = \text{MIN}((w_{jm})^r + (w^i_{mk})^r)^{1/r}$, for $1 \le m \le n$.
2. Compute D^i, as follows: $d^i_{jk} = \text{MIN}(w^1_{jk}, \ldots, w^i_{jk})$, for $j \ne k$.
3. Iterate until W^q and D^q are computed.
4. Compare W^1 and D^q: all the links having the same values in these two matrices will belong to the final PFNET.

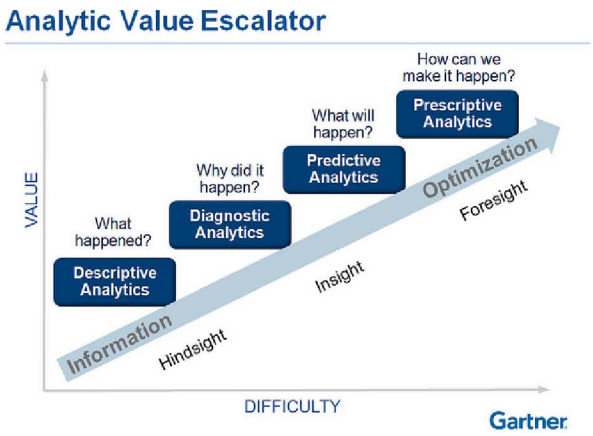
〈그림 7〉 패스파인더 네트워크 척도 알고리즘(Quirin et al., 2008)

가는 최소비용 값을 저장하며 W^1 초기가중치 값으로 W_{jk}^{i-1}을 재귀적으로 계산하여 산출할 수 있다. D_{jk}^{i}는 i개 이하의 링크로 구성되는 경로들을 따라 노드 j에서 k로 가는 최소비용 값을 저장하며, 이 매트릭스는 모든 W_{jk}^{1}, ..., W_{jk}^{i}에 대해 재귀적으로 계산하여 산출된다. 최종단계에서 W^1와 D^q의 두 매트릭스를 비교하여 같은 값을 갖는 모든 링크를 추출함으로써 PFNet을 생성한다.

PFNet 알고리즘은 구현이 어렵고 계산의 복잡도가 높다는 것이 단점으로 알려져 있는데, 이를 해결하기 위해 최소신장트리(Minimum Spanning Tree: MST) 방법을 기반으로 엄격한 조건인 $r=\infty$, $q=n-1$ 조건 상에서 빠르게 구현할 수 있는 MST-PF 알고리즘이 제안된 바 있으며(Quirin et al., 2008), PFNet을 구현하면서 동시에 군집분석을 수행하는 알고리즘인 PNNC(Parallel Nearest Neighbor Clustering)가 제안되기도 하였다(이재윤, 2006). 본 연구를 수행하기 위해 MST-PF 알고리즘을 기반으로 한 PFNet 생성 소프트웨어를 직접 개발하였으며, 네트워크 시각화를 위해 네트워크 분석을 위한 공개 소프트웨어로 유명한 Pajek (http://mrvar.fdv.uni-lj.si/pajek/)을 사용하였다.

2.3 추천의 근거와 선택의 옵션을 제공하는 처방적 분석 기법

세계적인 기업 환경 분석 그룹인 가트너(Gartner)는 〈그림 8〉에서 보는 바와 같이 분석 기술의 발전 단계를 크게 4가지로 구분하고 있다. 첫 번째 단계는 기술적 분석(descriptive analytics)으로 회귀적인(retrospective) 관점에서 데이터에 쌓여있는 현상을 파악하는 가장 기초적인 분석 방법을 사용하는 것이다(*What happened?*라는 질문에 대한 답을 제공한다). 두 번째 단계인 진단형 분석(diagnostic analytics) 방법에서는 왜 발생했는가하는 데이터 기반의 진단이 포함되며 분석적인 관점의 접근을 강조

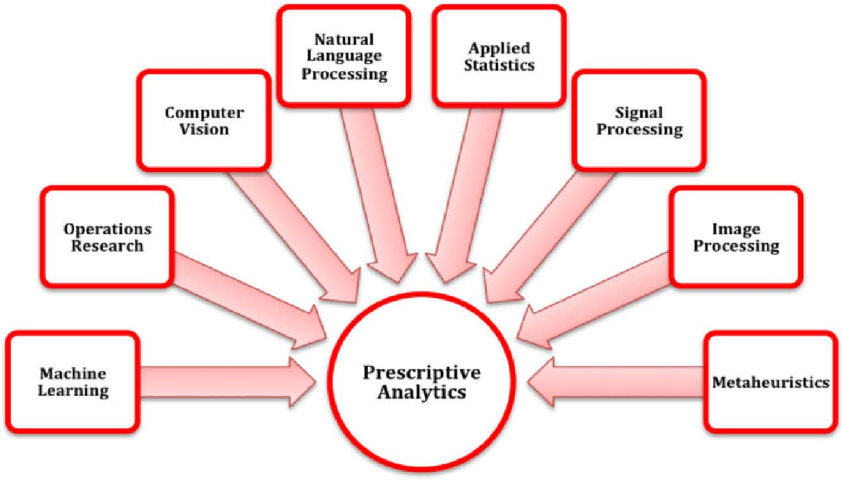


〈그림 8〉 분석 기술의 발전 단계와 주요 특징
(Gartner, https://www.gartner.com)

하기 시작하였다(*Why did it happen?*이라는 질문에 대한 답을 제공한다). 세 번째 단계인 예측적 분석 기법(predictive analytics)은 기존의 데이터로부터 판단하건데 앞으로 어떤 추세가 나타날 것인가를 분석하는 방법론을 제공한다(*What will happen?*이라는 질문에 대한 답을 제공한다). 마지막 단계인 처방적 분석 기법(prescriptive analytics)은 기존의 방법론 들을 융합하여 개념적으로 가장 진화된 분석 방법들을 제공한다. 이러한 처방적 분석 기법의 목표는 시스템이 제공하는 최적화된 옵션들 중에서 어떤 것을 선택하고 그 결과가 어떻게 될 것인가를 예상할 수 있는 과학적인 근거를 제공하는 데 있다(*How can we make it happen?*이라는 질문에 대한 답을 제공한다).

기술적인 면에서 볼 때, 처방적 분석 방법론은 기존의 다양한 기술적 요소를 융합적으로 사용하며 기존의 방법론들에 비해 최적화와 시뮬레이션에 초점을 두고 있다. 기계학습, 운용과학(수학과 통계를 기반으로 최적의 시스템을 개발), 컴퓨터비전, 자연어처리, 응용통계, 신호처리, 이미지처리, 메타휴리스틱스(최적화 알고리즘을 개념적으로 해석적용) 등의 복합적 기술요소를 효과적으로 융합하고 적용을 하는 융합기술(convergence technology)의 개념으로 볼 수 있다(〈그림 9〉 참조). 특히 기계학습 기법은 예측적 분석과 처방적 분석과 같은 분석적 시스템의 결과 도출을 위한 가장 중요한 기반 기술 중 하나이다(Bertsimas & Kallus, 2015). 본 연구에서는 자연어 처리와 텍스트 마이닝 기법들을 활용하여 최종 시스템에 기계학습과 시각화 기법을 적용한다.

처방적 분석 개념을 도입한 지능형 시스템의 연구 개발 사례로, Song et al.(2013)은 처방적 분석 시스템의 유용성을 소개하면서 다양한 정보원으로부터 수집된 정보를 기반으로 다수의 복합적 지표를 산출하여 연구자의 대표 활동 유형과 비교 활동 성향을 파악하였다. 이 시스



〈그림 9〉 처방적 분석 기법을 구현하기 위한 다양한 기술 분야
(Wikipedia: Prescriptive analytics, 2017)

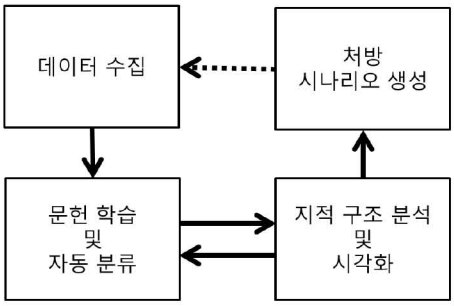
템을 통해 연구자는 본인의 현재 역량을 분석하고 최종 성과 목표에 도달하기 위한 몇 가지 처방적 지시를 얻을 수 있음을 보여주었다. 김장원, 황명권, 송사광, 김진형, 정도헌, 정한민(2014)은 처방적 분석 기법의 개념을 도입하여 빅데이터 환경에서 대량으로 생성되는 연구자 정보를 수집하고 분석하였다. NDSL, DBLP, Linked-In으로부터 연구자 정보를 수집하고 이를 학술활동, 산업활동, 경력활동의 세가지 카테고리로 분류한 후 행동 유형을 상세 분석함으로써 연구자의 학술 활동 히스토리를 추적하는 통합 분석 시스템을 제안하였다.

이어 3장에서는 문헌 분류를 기반으로 하는 응용 시스템을 최적화하고 효율적으로 운영하는 방안을 처방적 분석의 관점에서 다루고, 이를 구현하는 일련의 과정을 설명한다. 기계학습 기법과 네트워크 분석 및 시각화 기법을 설명하고 이 두 가지 방법을 결합하여 처방적 분석의 개념으로 발전시키는 방안을 제시한다.

3. 처방적 분석 개념을 적용한 문헌 분류 시스템 설계

3.1 전체 시스템 개요

본 연구에서 제안하는 전체 시스템은 〈그림 10〉과 같이 크게 4개의 컴포넌트로 이루어져 있다. 우선, 데이터 수집부에서는 대량의 학술 논문을 수집하고 저장한다. 자동 분류 성능 평가를 위해 모든 논문에는 분류정보가 입력되어 있으며 일부는 학습용 데이터 집합으로 사용하고 나머지는 성능 평가용 테스트 데이터 집합으로 사용한다. 그 다음 단계는 문헌 분류 단계이다. 수집된 대용량 문헌을 자동 분류기를 통해 학습 및 분류를 실시한다. 본 연구에서 활용하는 FVC는 확률적 분류기 유형으로, 이 단계의 결과로 나타나는 분류 결과 정보는 *ST-30=(ST-30:0.4, ST-28:0.3, ST-25:0.1, ST-26:0.1, ...)*과 같이 주제 범주 간의 확률적인 유사도 형태로 나타난다. 따라서 이 확률 데이터를 활용하면 시스템이 가장 모호하게 분류한 문헌 데이터들을 수집할 수 있다. 이는 앞서 소개한 능동적 학습을 통한 문헌 데이터의 품질 개선 시스템으로 쉽게 확장 발전시킬 수 있다. 단, 능동적 학습 과정은 본 연구의 범위에서 벗어나므로 자동 분류 실험에서는 다루지 않는다. 다음 단계의 시스템 구성요소는 지적 구조 분석을 통한 시각화 단계이다. 자동 분류시 생성된 오류 로그를 이용하면 유사계수(similarity coefficient)를 사용하여 학문 분야 간 유사도를 측정하는 방법과 동일하게 학문 분야의 전역적 네트워크를 생성할 수 있다. 자동 분류의 결과를 이용해 네트워크 구조 분석을 수행하기 때문에 전체 시스템은 각 컴포넌트들이 유기적으로 연동되도록 설계되어 있다. 마지막 단계는 시나리오 생성부이다. 실험 결과를 해석함으로써 처방적 분석 시스템에서 제공할 수 있는 몇 가지 분류 체계 조정 시나리오를 도출하는 단계이다. 특히 2단계와 3단계인 자동 분류와 시각화는 빠르게 상호 작용을 하며 최적화를 위한 반복 계산을 수행한다. 특히 시스템을 생명력 있게 지속적으로 운영하기 위해서는 최종 단계에서 첫 단계로 자연스럽게 순환되도록 피드백 과정을 두는 것이 바람직하다. 다음 장부터 네 개의 컴포넌트로 구성된 시스템 설계 방안을 단계별로 설명한다.



〈그림 10〉 처방적 분석 시스템 구성 컴포넌트

3.2 데이터 수집과 분류 체계

자동 분류 실험을 위해 해외 학술논문 데이터를 수집하였다. 한국과학기술정보연구원의 과학기술 정보 오픈서비스인 NOS(NDSL Open Service, http://nos.ndsl.kr)를 이용하면 오픈 API 방식으로 기 구축된 대량의 학술정보를 검색하고 메타데이터를 다운 받을 수 있다. 수집기는 모두 python 언어를 이용해 직접 개발하였다. 수집한 해외 학술 논문 정보에는 저널 단위로 DDC(Dewey Decimal Classification) 코드가 상세하게 부여되어 있으며, 특히 과학기술 중심의 데이터 셋은 매우 상세한 수준이기 때문에 유사한 학문 영역은 1차 그룹핑을 하여 학습 데이터 셋의 규모가 일정한 양이 되도록 기준 분류 체계의 조정을 실시하였다. 이를 바탕으로 자동 분류를 수행하기에 앞서 전체 실험대상 논문에 대해 변환표를 이용해 42개의 새로운 범주체계로 일괄 변환처리 하였다(〈표 1〉 참조). 전체 수집된 학술논문은 약 73만 건으로 학문 주제 분야별 구축 양과 비율은 〈표 2〉와 같다.

〈표 1〉 데이터 분류를 위한 42개 주제코드 변환표

학문 주제 분야	DDC 분류	주제 코드
전산, 정보	003.* ~ 02*	ST-01
기타총류	000.* ~ 002.* / 03* ~ 09*	ST-02
철학	10* ~ 14* / 16* ~ 19*	ST-03
심리학	15*	ST-04
종교	2**	ST-05
사회과학일반	30*, 31* / 36*, 39*	ST-06
정치	32*	ST-07
경제	33*	ST-08
법률	34*	ST-09
행정	35*	ST-10
교육	37*	ST-11
무역	38*	ST-12
언어	4**	ST-13
자연과학일반	50*	ST-14
수학	51*	ST-15
천문	52*	ST-16
물리	53*	ST-17
화학	54*	ST-18
지학, 지질	55* ~ 56*	ST-19
생명과학	57*	ST-20
식물	58*	ST-21
동물	59*	ST-22
기술과학일반	60*	ST-23
의학일반	610.*	ST-24
기초의학	611.* ~ 612.*	ST-25
사회의학	613.* ~ 614.*	ST-26
약학, 치료학	615.*	ST-27
내과학	616.* , 619.*	ST-28
외과학	617.*	ST-29
산부인과	618.*	ST-30
공학일반	620.*, 629.*	ST-31
응용물리	621.*	ST-32
건설, 건축, 토목	622.* ~ 627.* / 69*	ST-33
도시, 환경	628.*	ST-34
농업	63*	ST-35
가정, 가사	64*	ST-36
경영	65*	ST-37
화학공학	66*	ST-38
제조	67* ~ 68*	ST-39
예술	7**	ST-40
문학	8**	ST-41
역사, 지리	9**	ST-42

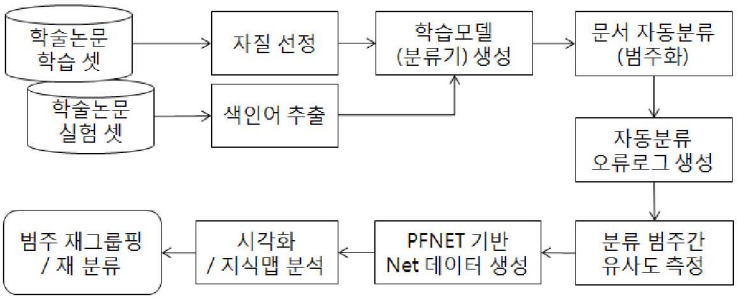
3.3 대용량 문헌의 자동 분류를 통한 학문간 연관 정보 생성

본 연구에서 제안하는 처방적 분석 시스템을 위한 전체 실험 과정은 〈그림 11〉과 같다. 첫 번째는 기계 학습 기반의 자동 분류 과정으로써, 논문으로부터 적절한 자질선정 과정을 거쳐 학습모델을 생성한 후 자동 분류를 수행한다. 두 번째는 자동 분류 결과로 생성된 오류 로그를 이용해 학문의 주제 분야 간 연관 정보를 생성한다. 세 번째는 네트워크 생성 기법과 시각화 도구를 이용해 전체 주제 분야의 지식 맵을 작성한다. 마지막으로 지식 맵을 해석하고 이를 근거로 범주를 재설정하여 자동 분류를 재수행한 후 결과를 확인한다.

일반적으로 학술 논문 등의 자동 분류를 수행하는 경우에 주제 분야의 특성, 즉 분야에서 사용하는 용어의 전문성 정도의 차이로 인해 성능의 편차가 나타나는 경향이 있다. 인문, 사회과학 분야 보다는 과학기술 분야가 자동 분류의 성능이 우수하게 나타나며, 여러 분야가 연계, 융합된 응용 연구 분야 보다는 순수과학 분야가 쉽게 식별이 되는 경향이 있다. 용어의 전문성이 높은 의학 분야의 논문은 다른 분야에 비해 분류 정확률이 매우 높게 나타나는 경향을 보여주었다(정도헌 외, 2007).

본 연구를 위해 수집한 대용량 해외학술논문 데이터를 FVC 방식의 분류기로 학습하고 자동 분류한 결과는 〈표 2〉와 같다. 학습과 실험 셋은 8:2의 비율이며, 교차검증 방법인 5-fold cross validation을 이용하여 평균값을 최종 성능 값으로 산출하였다. 실험 결과, 학문 영역 간에 매우 심한 성능 편차가 나타났는데, F1 척도를 기준으로 화학 분야(60.92%)가 가장 높은 성능 수준이었다. 전체 데이터 셋의 마이크로 F1 점수(micro averaged F1 score)는 44.83%, 매크로 F1 점수(macro averaged F1 score)는 34.39%로, 초기 데이터의 자동분류 결과는 전체적으로 낮은 성능을 보여주었다.

일단 자원에 부여된 분류 코드 값은 명확한 오류가 발견되지 않는 한 수정하는 경우는 드물다. 현재 사용하는 주제 분류 체계가 일관성이 낮고 실제 데이터의 특성을 잘 반영하지 못하는 상태라면, 데이터 현황을 점검하여 기존의 주제 분류 체계를 재설정할 필요가 있을 것이다. 이러한 개선 프로세스가 없이는 낮은 서비스 품질이 지속될 가능성이 높기 때문에 기계 학습을 통한 분석 결과가 데이터의 품질 보정에



〈그림 11〉 지적 구조 분석을 위한 실험 개요도

활용되도록 하는 것이 매우 중요하다. 따라서, 시스템을 설계할 때 처방적 분석 개념을 도입함으로써 데이터의 품질 개선을 위한 피드백 과정을 실행할 수 있다. 본 연구에서는 분류 체계를 합리적으로 재설정하기 위하여 주제 범주 간의 연관성을 측정하고 지적 구조를 파악하여 의사결정에 필요한 분석 데이터를 산출하는 과정을 제시한다.

〈표 2〉 42범주 체계의 주제 분야별 자동 분류 수행 결과

주제 분야 /주제 코드	문서 수	F1
전산, 정보 /ST-01	11,225	0.3478
기타총류 /ST-02	7,931	0.1393
철학 /ST-03	1,857	0.1689
심리학 /ST-04	6,308	0.2940
종교 /ST-05	2,300	0.2379
사회과학일반 /ST-06	28,094	0.3960
정치 /ST-07	6,572	0.2633
경제 /ST-08	20,742	0.4308
법률 /ST-09	2,987	0.1424
행정 /ST-10	2,593	0.2508
교육 /ST-11	9,828	0.4813
무역 /ST-12	2,261	0.1995
언어 /ST-13	1,291	0.2357
자연과학일반 /ST-14	12,712	0.0429
수학 /ST-15	22,538	0.5576
천문 /ST-16	3,297	0.4330
물리 /ST-17	56,620	0.4620
화학 /ST-18	56,303	0.6092
지학, 지질 /ST-19	22,059	0.5710
생명과학 /ST-20	82,876	0.5167
식물 /ST-21	32,181	0.5888
동물 /ST-22	22,194	0.3416
기술과학일반 /ST-23	392	0.1202
의학일반 /ST-24	10,644	0.1458
기초의학 /ST-25	10,016	0.1062
사회의학 /ST-26	11,585	0.2336
약학, 치료학 /ST-27	18,596	0.3201
내과학 /ST-28	93,436	0.5940
외과학 /ST-29	30,565	0.4683
산부인과 /ST-30	8,235	0.3537
공학일반 /ST-31	25,907	0.3732
응용물리 /ST-32	39,897	0.3774
건설, 건축, 토목 /ST-33	5,345	0.3202
도시, 환경 /ST-34	3,721	0.3179
농업 /ST-35	13,086	0.4068
가정, 가사 /ST-36	842	0.2076
경영 /ST-37	7,776	0.3161
화학공학 /ST-38	21,122	0.3196
제조 /ST-39	6,467	0.2619
예술 /ST-40	4,777	0.1323
문학 /ST-41	507	0.0678
역사, 지리 /ST-42	4,218	0.2281
총 합	731,903	
매크로 평가 (macro averaged eval.)		0.3439
마이크로 평가 (micro averaged eval.)		0.4483

※ 성능 상위 3개 분야를 음영 표시함

자동화된 일련의 과정을 갖는 처방적 분석 시스템을 설계하기 위해 자동 주제 분류의 결과로 생성된 오분류 로그 정보를 이용하여 주요 학문영역 간의 지적 구조를 파악하는 방법을 제시한다. 계량정보학은 주로 저자 또는 문헌 간 인용 정보, 동시발생 용어(co-word), 학문 주제범주 및 웹 링크 등의 주요 자질 요소 간의 연관성을 측정하고 시각화하여 최종 분석을 하는 일련의 과정을 수행한다. 이에 반해, 본 연구는 정보 서비스를 위한 문헌의 자동 분류 과정에서 발생한 오분류의 빈도 정보를 이용하여 학문의 주요범주 간 연관성을 측정한다는 점에서 차이가 존재한다. 동시 발생 요소 간의 연관 빈도행렬에 대해 상관 정도를 측정하는 과정을, 자동 분류의 실패로 판정된 오류 빈도 정보를 이용하여 주제 간의 유사도를 측정하는

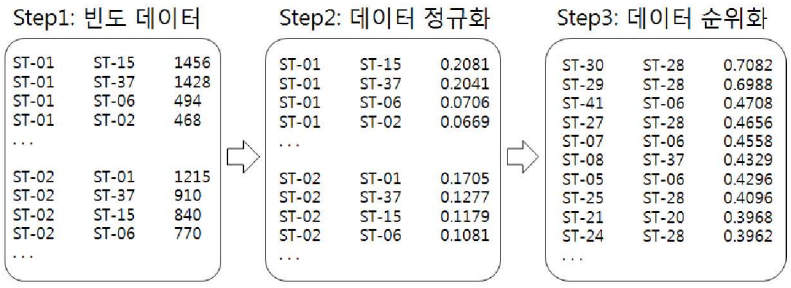
과정으로 대체하여 이를 직접 학문분야의 지적 구조 분석용 데이터로 활용한다. 이러한 방식으로 데이터를 분류하고 그 결과를 시각화 하여 구조를 검증하는 과정을 이음새없이 연결함으로써 효율적으로 동작하는 처방적 분석 시스템을 개발할 수 있다.

확률 모델을 기반으로 하는 FVC 기반 분류 기법은 산출 값을 모든 주제 범주에 대한 유사도의 벡터 형태로 갖고 있게 된다(Jeong et al., 2012). 따라서 벡터 정보로부터 분야별 가중치를 측정하고 정규화한 후 이를 순위화하여 학문 주제 분야 간 최종 연관 정보를 산출할 수 있다. 마지막으로 최종 산출된 연관 정보를 이용해 패스파인더 네트워크 스케일링 알고리즘에 기반한 네트워크 시각화 기법을 통해 전체 학문의 지적 구조를 파악한다.

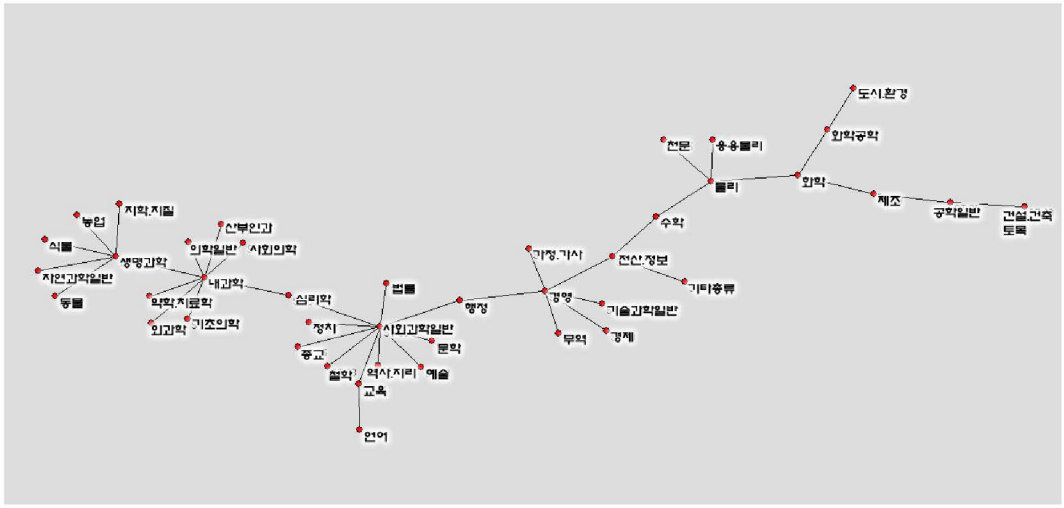
학문 영역의 지적 구조 분석을 위한 주제 범주간 연관 데이터 생성 과정은 다음과 같다. 일반적으로 자동 분류를 수행한 결과로써 정분류와 오분류 값이 산출되게 되며, 능동적 학습 기법 등을 이용해 오류에 대한 정보는 분류성능 향상을 위해 재사용될 수 있다(McCallum & Nigam, 1998; Settles, 2010). 본 연구에서는 오답으로 산출된 로그정보를 이용해 유사한 인접 학문을 연관 정보로 생성한 후 지적구조 분석을 위한 정보로 활용한다. 자동 분류 프로세스를 통해 생성된 오류 로그 데이터는 〈그림 12〉와 같이 초기 빈도 데이터 형태로 간단하게 산출될 수 있다(step1). 산출된 데이터 예시인 Step1의 첫 번째 '*ST-01 ST-15 1456*'은 정답 범주는 ST-01이지만 ST-15라고 판단한 오류 문헌이 1,456건 임을 나타낸다. 이 때, 각 주제 범주별 전체 문헌의 건수가 상이하므로 범주별 문헌수의 전체로 각 빈도를 나눈 값으로 정규화를 거친 후(step2), 마지막 단계에서 유사도 순위 기준으로 재 정렬을 수행한다. 순위화된 두 주제 범주간의 연관 데이터는 네트워크 기반의 분석 알고리즘에 적용하기 위한 기본 입력 포맷이 된다(step3).

3.4 학문분야의 지적 구조 분석 및 네트워크 시각화

PFNet 알고리즘을 통해 그려진 주제 범주간 지적 구조는 〈그림 5〉와 같다. 네트워크 시각화를 통해 분석한 결과, 학문 주제영역에 대한 지적 구조의 기본 골격은 "**[생명과학] – [의학] – [심리학, 사회과학, 행정학] – [경영, 경제] –**



〈그림 12〉 자동분류 오류 정보의 연관 데이터 변환 생성 과정



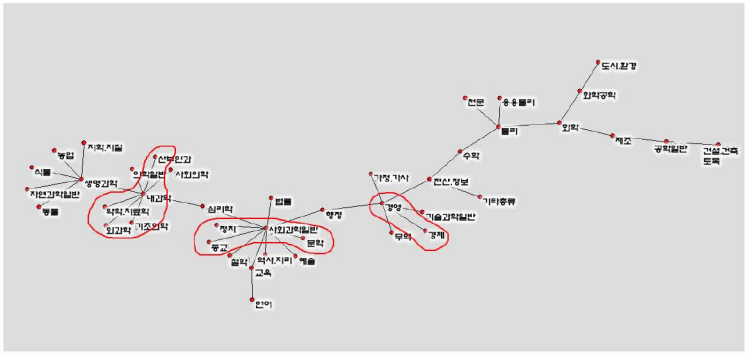
〈그림 13〉 PFNet 알고리즘에 기반한 42개 주제 범주의 지적 구조 시각화
(Pajek의 Kamada-Kawai 드로잉 알고리즘을 사용함)

[전산, 정보, 수학] – [물리학] – [화학] – [건설, 건축, 토목]"으로 이루어져 있음을 확인할 수 있다. 주제영역 간의 연결매개, 즉 허브 역할을 하는 세 개의 주제 범주가 확연히 드러나는데, 첫 번째는 '심리학'이다. 심리학은 의학과 인문, 사회과학 그룹 사이에 존재하고 있으며, 정신의학 분야와 사회심리학 등의 다양한 응용분야와 관련되어 있는 학문적 특성을 보이고 있다. 동시에 인지과학 영역의 주된 학문이 철학, 심리학, 인공지능, 신경과학 등임을 고려해 볼 때, 〈그림 13〉과 같이 인문, 사회계열에 다소 가깝지만 의학과 상호 연결하는 모습은 타당한 학문의 흐름으로 볼 수 있다. 두 번째 허브는 '행정학'으로 인문, 사회과학 그룹과 경영, 경제 그룹을 상호 연결하고 있다. 그래프 상으로는 정치, 법률 분야보다 행정 분야가 경제관련 분야에 더 큰 영향을 주는 것으로 해석할 수 있다. 세 번째 허브는 '수학'으로 전산, 정보와 함께 그룹핑을 하는 경우가 많은 분야이지만 분명하게 응용학문인 전산영역과 기초학문인 물리영역을 연결하고 있음을 확인할 수 있다. 이와 같이 PFNet을 이용한 42개 주제 분야의 지식 맵을 통해 학문의 전역적인 위상 구조와 함께 데이터의 의미적인 관점에서 연관성이 높은 인접 주제 분야를 쉽게 식별할 수 있다.

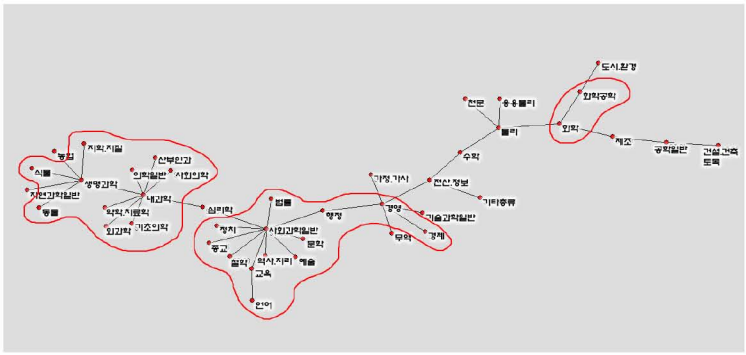
자동 분류 결과로 얻어진 오분류 로그 정보는 가장 유사한 두 개의 학문 주제 분야 간의 링크 정보를 표현하고 있음을 앞서 언급하였다. 학문 간 연결강도인 링크 유사도(cosine 유사계수를 사용)를 기준으로 임계치를 변화시키면서 단계별로 생성되는 학문분야의 그룹을 비교한 결과는 다음과 같다. 초기에는 가장 강하게 연결되는 세 개의 그룹이 나타나는데 의학의 주요 분야와 인문·사회과학의 주요 분야, 그리고 경영, 경제 그룹이다(〈그림 14〉 참조). 이 단계를 조금 더 확장하여 표현하면 〈그림 15〉와 같이 나타나는데 이 단계에서의 주요 특징은 의학과 생명과학 영역이 매우 밀접하게 연결이 되는 현상과 인

문·사회과학의 전 영역과 경영, 경제 분야가 하나의 그룹이 되는 현상, 새롭게 화학, 화학공업의 그룹이 생겨난 현상 등이다.

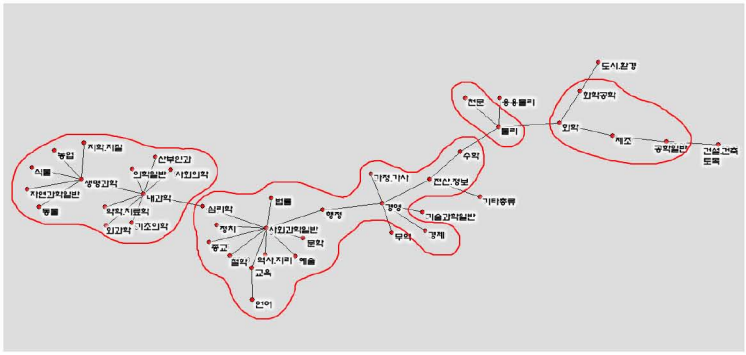
유사도의 임계치를 더욱 낮추면 관련분야의 유사 분야의 영역이 더욱 커져서 크게 네 개의 영역으로 나뉘게 되며(〈그림 16〉 참조), 유사도 임계치가 0.18 이하에서 응용물리와 기술과학일반이 제외된 나머지가 모두 한 개의 클러스터로 형성이 됨을 알 수 있다(〈그림 17〉 참조). 기술과학일반은 본 실험에서 실험대상 문헌의 수가 매우 적고 범주특성상 속성이 모호한 것으로 이해되지만, 응용물리학의 영역이 초기 클러스터에서 물리학에 편입되지 않은 것은 다소 의외였다. 그러나 응용물리학(applied physics)은 연결강도는 약하지만 유일하게 물리학과 직접 연결되어 있는 종단 노드로써 명확한 위상을 가지고 있으며, 속성상 매우 강한 학제적 성격을 갖고 있음을 감안할 때(Applied physics, 2017), 결과적으로 PFNet 알고리즘은 전역적인 지적 구조의 네트워크 상에서 적절한 학문간 위상구조를 보여주는 데 효과적이었다.



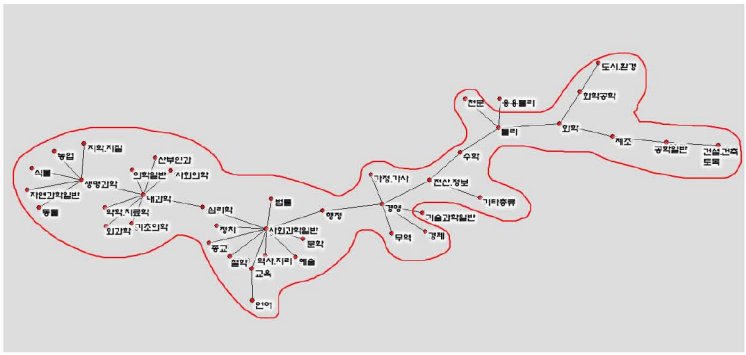
〈그림 14〉 의학, 인문사회과학, 경영경제 분야 그룹이 두드러짐
(코사인 유사도 0.4 이상 주제 분야 클러스터링)



〈그림 15〉 의학과 생명과학, 인문사회와 경영경제가 융합, 화학과 화학공학 분야 그룹이 등장(코사인 유사도 0.25 이상 주제 분야 클러스터링)



〈그림 16〉 영역이 확장되며 전체가 네 개의 유사 그룹으로 형성됨
(코사인 유사도 0.2 이상 주제 분야 클러스터링)



〈그림 17〉 일부를 제외하고 전 분야가 하나의 그룹으로 형성
(코사인 유사도 0.18 이상 주제 분야 클러스터링)

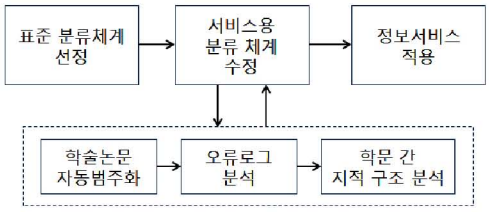
4. 처방적 분석 개념을 적용한 시나리오 도출

분석 기술의 발전은 축적된 데이터를 일목요연하게 보여줌으로써 현상을 파악하는 기술적 분석론(descriptive analytics)과 진단적 분석론(diagnostic analytics), 추세를 통해 미래를 추정하는 예측적 분석론(predictive analytics)을 거쳐 최근에는 최적화 기법을 통해 선택 가능한 여러 시나리오를 시스템이 제시할 수 있다는 처방적 분석 방법론(prescriptive analytics)에까지 이르렀다. 처방적 분석 방법의 개념은 점차 다양한 학문과 서비스, 응용 시스템에 활용되며 지속적으로 발전하고 있다.

앞서 자동 분류의 오류로그를 이용하는 계량정보 분석적 방법이 주요 학문 영역의 전체적인 구조 파악에 유용하며, 학문분야 간의 연관정도를 나타내는 지적 구조를 고려하여 분류체계를 재설정하는 근거가 될 수 있음을 확인하였다. 본 장에서는 기계 학습과 계량정보 분석 방법론을 융합하여 시나리오를 도출함으로써 처방적 분석 시스템으로 발전시키는 방안에 대해 살펴

본다.

처방적 분석 개념을 적용함으로써 기존의 분류 서비스 시스템을 의사결정 지원 시스템(decision support system: DSS)으로 재설계하여 활용할 수 있다. 〈그림 18〉에서 보는 바와 같이 표준 분류 체계를 선정하고 서비스하는 과정에서 서비스 개선점을 도출하는 기능을 탑재할 수 있다. 정보 검색 및 분석 서비스를 고려하여 범주를 재설정하는 과정에서 변경의 근거를 토대로 분류 체계를 효과적으로 재설정함으로써 최종 서비스의 지속적인 성능 향상을 꾀할 수 있다.



〈그림 18〉 처방적 분석 시스템에 적용된 지적 구조 분석 프로세스 위치

앞선 자동분류 실험의 결과로 F1 값이 분야별로 차이가 심하게 나타남을 확인하였다. 특히 적은 문헌 수를 가지는 학문 분야의 낮은 분류 성능으로 인해 매크로 F1 값(macro averaged F1 score)이 마이크로 F1(micro averaged F1 score)보다 저조함을 확인하였다. 이에 매크로 평가의 측면에서는 전체 시스템의 성능 저하를 해결하기 위해 기존 주제 분류체계의 구조 조정이 필요할 것으로 보인다. 주제 분류체계의 조정은 전문가에 의한 수작업으로 이루어지지만 조정을 위한 근거 자료로써 학문의 전역적 지적 구조 분석 작업이 수행되었음을 앞서 설명하였다. 처방적 분석 시스템을 통해 도출된 특정 패턴들을 기반으로, 시스템 분석가는 시스템에 적용가능한 몇 가지 후보 시나리오를 생성할 수 있다. 분석 결과를 기반으로 42개의 분류체계를 최종 29개의 주제 분류 체계로 재조정한 단계별 시나리오는 다음과 같다.

첫 번째 생성 규칙(production rule)은 앞서 클러스터링 기법을 이용해 시뮬레이션한 바와 같이 지식 맵 상에서 범주 간 유사도가 높은 학문 분야를 그룹핑하는 것이다. 이 결과로 ST-24, ST-25, ST-26, ST-27, ST-28, ST-29, ST-30 분야가 통합되어 의・약학(STMED라 명명함) 카테고리를 생성하였다. 같은 방식으로 ST-08, ST-37를 경영・경제(STECO*)로, ST-18, ST-38을 화학・화학공업(STCHE)으로 그룹핑하였다. 두 번째 생성 규칙은 상호관계의 명시적 설명이 가능한 주제 분야를 고려해 그룹핑하는 것이다. 이 과정을 수행하여 ST-01, ST-15 분야는 전산・정보・수학을 의미하는 STINF라는 코드로 통합되었다. 또한 앞서 논의 되었던 물리와 응용물리 분야인 ST-17, ST-32는 STPHI로, ST-11, ST-13은 교육・언어(STEDU) 그룹으로 생성하였다. 세 번째 규칙은 문서수가 적고 분류 성능이 낮으며, 인접한 유사 주제 분야를 그룹핑 하는 단계이다. 앞서 1단계에서 생성했던 STECO* 그룹에 무역(ST-12)을 포함하여 최종 그룹 STECO를 생성하였다. 이 경우에는 규칙1을 적용한 후 다시 규칙3을 적용한 경우이다. 마지막으로 문학과 예술을 다루는 ST-40, ST-41의 두 분야를 통합하여 START 코드를 부여하였다.

생성 시나리오에 기반하여 분류 체계를 변경한 후 재수행한 자동 분류 성능의 결과 비교 그래프는 〈그림 19〉와 같다. 마이크로 F1 값이

〈생성 규칙(production rule)〉

- Rule 1: 지식 맵 상의 주제 범주간 유사도 고려해 그룹핑
- Rule 2: 상호관계의 명시적 설명이 가능한 주제 분야를 고려해 그룹핑
- Rule 3: 문서수가 적고 분류 성능이 낮으며, 인접한 유사 주제 분야를 그룹핑

〈시나리오 1: Rule 1 적용〉
→ STMED(의 · 약학): ST-24, ST-25, ST-26, ST-27, ST-28, ST-29, ST-30
→ STECO*(경영 · 경제): ST-08, ST-37
→ STCHE(화학 · 화학공업): ST-18, ST-38

〈시나리오 2: Rule 2 적용〉
→ STINF(전산 · 정보 · 수학): ST-01, ST-15
→ STPHI(물리 · 응용물리): ST-17, ST-32
→ STEDU(교육 · 언어): ST-11, ST-13

〈시나리오 3: Rule 3 적용〉
→ STECO(경영 · 경제 · 무역): STECO*, ST-12
→ START(문학 · 예술): ST-40, ST-41

〈그림 19〉 범주 체계와 학문 계열에 따른 자동 분류 성능 비교

매크로 F1 값보다 전체적으로 모두 우수한 것으로 보아 분야별 성능의 편차가 매크로 평가에 영향을 준 것으로 보인다. 전체적으로 볼 때, 과학기술 분야에 비해 인문 · 사회 계열의 문서 분류 성능이 상당히 낮은 결과를 보이고 있다. 42범주와 29범주 기준을 비교해 보면, 전체 데이터를 그룹핑하여 재학습한 결과로 인해, 매크로 평가와 마이크로 평가에서 모두 29범부의 분류 성능이 향상되었다. 유사한 분야를 묶어 재학습하였기 때문에 기대했던 시나리오 반영의 효과라 볼 수 있다. 세부 주제 분야별 성능을 보면, 42범주 체계에서는 화학 분야와 내과학

분야가 각각 60.92%, 59.4%의 최고 분류 성능을 보이며(〈표 2〉 참조), 29범주 체계에서는 의・약학 분야와 화학・화학공업 분야가 각각 76.84%, 62.23%의 최고 성능을 보이고 있다(〈표 3〉 참조). 42범주와 29범주 체계 모두에서 일반 영역인 '사회과학일반, 자연과학일반, 그리고 기술과학일반' 범주는 전체적인 성능 저하의 원인이 되고 있으며, 인문・사회과학 영역의 주제 범주 중 많은 경우가 10%대의 매우 낮은 성능을 보이고 있음을 감안할 때, 처방적 분석 시스템의 진단 결과를 참고해서 데이터 수준의 품질 개선, 분류 체계 재조정 등은 계속 이루어져야 할 것으로 보인다.

〈표 3〉 축소된 29범주 체계의 주제 분야별 자동 분류 수행 결과

주제 분야 / 주제코드	학습문서 수	F1
전산・정보・수학 /STINF	33,763	0.5465
기타총류 /ST-02	7,931	0.1393
철학 /ST-03	1,857	0.1689
심리학 /ST-04	6,308	0.2940
종교 /ST-05	2,300	0.2379
사회과학일반 /ST-06	28,094	0.3960
정치 /ST-07	6,572	0.2633
경영・경제・무역 /STECO	30,779	0.5333
법률 /ST-09	2,987	0.1424
행정 /ST-10	2,593	0.2508
교육・언어 /STEDU	11,119	0.4748
자연과학일반 /ST-14	12,712	0.0429
천문 /ST-16	3,297	0.4330
물리・응용물리 /ST-17	96,517	0.5425
화학・화학공업 /ST-18	77,425	0.6223
지학, 지질 /ST-19	22,059	0.5710
생명과학 /ST-20	82,876	0.5167
식물 /ST-21	32,181	0.5888
동물 /ST-22	22,194	0.3416
기술과학일반 /ST-23	392	0.1202
의・약학 /STMED	183,077	0.7684
공학일반 /ST-31	25,907	0.3732
건설, 건축, 토목 /ST-33	5,345	0.3202
도시, 환경 /ST-34	3,721	0.3179
농업 /ST-35	13,086	0.4068
가정, 가사 /ST-36	842	0.2076
제조 /ST-39	6,467	0.2619
문학・예술 /START	5,284	0.1140
역사, 지리 /ST-42	4,218	0.2281
총 합	731,903	
매크로 평가		0.3712
마이크로 평가		0.5485

※ 성능 상위 3개 분야를 음영 표시함

5. 결 론

본 연구는 새로운 분석 방법인 처방적 분석 기법의 개념을 도입하여 분석의 결과를 제시함과 동시에 최적화된 결과가 나오기까지의 과정을 제공함으로써 문헌 분류를 기반으로 하는 응용 시스템을 더욱 쉽게 최적화하고 효율적으로 운영하는 방안을 제시하고자 하였다. 구축된 학술 문헌의 데이터 품질 현황을 진단하고 개선하기 위한 시나리오를 도출하기 위한 일련의 실험을 실시하였다. 우선, 대용량의 학술문헌을 수집하고 자질값 투표방식의 분류기(FVC)를 이용하여 자동 분류를 실시하였다. 실시 과정에서 발생한 오분류 로그 데이터를 이용하여 학문 간의 연관성을 파악하는 데 사용하여 전체 분석 과정이 유기적으로 연동되도록 설계하였다. 학문영역 간의 전역적인 지적 구조를 시각화하고 해석하기 위한 방법으로 네트워크 스케일링 알고리즘의 하나인 패스파인더 네트워크(PFNet)을 이용하여 42개 주제 분야에 대한 지식 구조를 작성하고 해석하였다. 마지막으로 지식 구조의

해석 과정에서 발견된 패턴 분석을 통해 문헌 분류 시스템에 적용할 수 있는 몇 가지 최적화의 시나리오를 도출하였다.

본 연구의 의의는 다음과 같다. 첫째, 정보 서비스를 위해 실시하는 자동 분류의 과정에서 발생하는 오분류 확률 데이터를 이용하여 계량정보학적 분석을 수행할 수 있다. 자동 분류를 수행함과 동시에 처방적 분석 시스템으로 발전시키는 데 매우 유용한 지적 구조 정보를 추가로 획득할 수 있다는 장점을 갖는다. 둘째, 지적 구조 정보를 해석함으로써 현재 데이터의 현황 정보와 연관 관계를 명확히 파악할 수 있으며, 의사결정 과정인 정보의 재구조화 작업을 보다 효과적으로 할 수 있는 명확한 근거를 마련할 수 있다. 설계된 시스템은 구조적으로 매우 신속하게 재학습 과정을 수행할 수 있으므로 문헌 분류 수준을 다양하게 변화시키면서 전체 문헌 집단의 구조가 어떻게 변화하는 가에 대한 시각화 데이터를 빠르게 생성하여 제공할 수 있다. 셋째, 처방적 분석 기법의 개념을 도입한 분류 시스템은 최적화의 과정을 통해 다양한 선택지를 사용자에게 제공할 수 있다. 학문 주제 간 유사도와 내용적 연관성을 고려한 범주 재설정 과정을 통해 시스템의 추천 성능과 신뢰성의 향상을 기대할 수 있다.

연구의 한계점은 다음과 같다. 본 연구에서 분석 시스템의 유저 인터페이스 설계는 연구의 범위에 포함되지 않았다. 즉, 시스템 설계 방법에 대안 제안과 요소 기술을 다루고 있으므로 각 단위 모듈인 정보 수집기, 자동 분류기, 시각화 그래프 생성기, 시나리오 생성기 등의 각 단위 시스템이 완전히 통합되어 있지 않다. 그러나 각 단계의 데이터 입출력은 논리적으로 연결되어 수행되었다. 전체 자동분류의 성능이 매우 저조하게 나타나는 정보의 품질 문제와 관련하여, 본 연구는 처방적 시스템의 도입을 제안하는 범위에 국한되어 수행되었음을 밝힌다. 따라서 서비스 시스템의 분류 정보 품질 향상과 관련하여 중요하게 언급되었던, 개별 단위 문헌의 오분류 정보를 수정 반영하는 방안은 다루지 않았다. 이는 능동적 학습(active learning)을 다루는 추가 연구를 통해 보다 심층적으로 수행되어야 한다.

향후에는 능동적 학습기술을 기반으로 한 고품질 데이터의 효과적 생성 방안에 대한 연구와 함께 처방적 분석 시스템의 자동 시나리오 생성을 위한 패턴 감지에 관한 텍스트 마이닝 기법 연구가 수행되어야 할 것이다.

참 고 문 헌

김장원, 황명권, 송사광, 김진형, 정도헌, 정한민 (2014). 지시적 분석을 위한 연구자 활동 기반의 연구자 히스토리 추적 서비스. 정보과학회논문지: 컴퓨팅의 실제 및 레터, 20(6), 359-363.

이원구, 신성호, 김광영, 정도헌, 윤화묵, 성원경, 이민호 (2011). 상호운용적 분류체계 관리를 위한 반자동 분류체계 관리방안. 한국콘텐츠학회논문지, 11(12), 466-474.

https://doi.org/10.5392/jkca.2011.11.12.466

이재윤 (2005). 문서측 자질선정을 이용한 고속 문서분류기의 성능향상에 관한 연구. 정보관리연구, 36(4), 51-69. https://doi.org/10.1633/jim.2005.36.4.051

이재윤 (2006). 지적 구조 분석을 위한 새로운 클러스터링 기법에 관한 연구. 정보관리학회지, 23(4), 215-231. https://doi.org/10.3743/kosim.2006.23.4.215

정도헌 (2010). 최대 개념강도 인지기법을 이용한 데이터베이스 자동선택 방법에 관한 연구. 정보관리학회지, 27(3), 265-281. https://doi.org/10.3743/kosim.2010.27.3.265

정도헌, 김환민, 김혜선, 신기정 (2007). 과학기술 전문용어의 주제 분야별 전문성과 자동 분류 성공률 간의 연관성 비교. 제14회 한국정보관리학회 학술대회 논문집, 31-36.

Applied physics (2017). Wikipedia Retreived from http://en.wikipedia.org/wiki/Applied_physics

Åström, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990-2004. Journal of the American Society for Information Science and Technology, 58(7), 947-957. https://doi.org/10.1002/asi.20567

Bertsimas, D., & Kallus, N. (2015). From predictive to prescriptive analytics. arXiv preprint arXiv:1402.5481.

Chua, A. Y. K., & Yang, C. C. (2008). The shift towards multi-disciplinarity in information science. Journal of the American Society for Information Science and Technology, 59(13), 2156-2170. https://doi.org/10.1002/asi.20929

Gartner (n.d.). Retrieved from https://www.gartner.com

Glänzel, W., & Schilemmer, B. (2007). National research profiles in a changing Europe (1983-2003): An exploratory study of sectoral characteristics in the Triple Helix. Scientometrics, 70(2), 267-275. https://doi.org/10.1007/s11192-007-0203-8

Hou, H., Kretschmer, H., & Liu, Z. (2008). The structure of scientific collaboration networks in Scientometrics. Scientometrics, 75(2), 189-202. https://doi.org/10.1007/s11192-007-1771-3

Jeong, D. H., Kim, J., Hwang, M., Song, S. K., & Jung, H. (2012). Classification method by integrating feature property matrices for large scale data. International Conference on SMA 2012, Kunming, China.

Kim, Heejung, & Lee, Jae Yun (2008). Exploring the emerging intellectual structure of archival studies using text mining: 2001-2004. Journal of Information Science, 34(3), 356-369. https://doi.org/10.1177/0165551507086260

Ko, Youngjoong, & Seo, Jungyun (2004). Using the feature projection technique based on a normalized voting method for text classification. Information Processing and Management,

40(2), 191-208. https://doi.org/10.1016/s0306-4573(03)00029-3

Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. Journal of the American Society for Information Science and Technology, 59(12), 1973-1984. https://doi.org/10.1002/asi.20914

McCallum, A., & Nigam, K. (1998). Employing EM and pool-based active learning for text classification. Proceedings of the Fifteenth International Conference on Machine Learning (ICML) 1998, 350-358.

Morillo, F., Bordons, M., & Gomez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. Journal of the American Society for Information Science and Technology, 54(13), 1237-1249. https://doi.org/10.1002/asi.10326

Pajek. [Computer software]. Retrieved from http://mrvar.fdv.uni-lj.si/pajek/

Prescriptive analytics (2017). Wikipedia. Retrieved from http://en.wikipedia.org/wiki/Prescriptive_analytics

Quirin, A., Cordon, O., Guerrero-Bote, V. P., Vargas-Quesada, B., & Moya-Anegon, F. (2008). A quick MST-based algorithm to obtain pathfinder networks(∞, n-1). Journal of the American Society for Information Science and Technology, 59(12), 1912-1924. https://doi.org/10.1002/asi.20904

Schvaneveldt, R. W. (1990). Pathfinder associative networks: Studies in knowledge organization. Norwood, NJ: Ablex.

Settles, B. (2010). Active learning literature survey. Computer Sciences Technical Report, 1648, University of Wisconsin-Madison

Song, Sa-Kwang, Kim, Donald J., Hwang, Myunggwon, Kim, Jangwon, Jeong, Do-Heon, Lee, Seungwoo, … Sung, Wonkyung (2013). Prescriptive analytics system for improving research power. Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on Computational Science and Engineering. https://doi.org/10.1109/cse.2013.169

Tague-Sutcliffe, J. (1992). An introduction to informetrics. Information Processing & Management, 28(1), 1-3.

White, H. D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. Journal of the American Society for Information Science and Technology, 54(5), 423-434. https://doi.org/10.1002/asi.10228

Zhao, D., & Strotmann, A. (2007). Can citation analysis of web publications better detect research fronts? Journal of the American Society for Information Science and Technology, 58(9), 1285-1302. https://doi.org/10.1002/asi.20617

• **국문 참고문헌에 대한 영문 표기**
(English translation of references written in Korean)

Gim, Jangwon, Hwang, Myung-Gwon, Song, Sa-Kwang, Kim, Jinhyung, Jeong Do-Heon, & Jung, Hanmin (2014). Researcher history tracking service for prescriptive analytics based on researcher activities. Journal of KIISE: Computing Practices and Letters, 20(6), 359-363.

Jeong, Do-Heon (2010). A study on automatic database selection technique using the maximal concept strength recognition method. Journal of the Korean Society for Information Management, 27(3), 265-281. https://doi.org/10.3743/kosim.2010.27.3.265

Jeong, Do-Heon, Kim, Hwan-Min, Kim, Hye-Sun, & Shin, Ki-jeong (2007). The relationship between the specificity of S&T terms and auto-classification accuracy. Proceedings of the 14th Conference of Korean Society for Information Management, 31-36.

Lee, Jae-Yun (2005). Improving the performance of a fast text classifier with document-side feature selection. Journal of Information Management, 36(4), 51-69. https://doi.org/10.1633/jim.2005.36.4.051

Lee, Jae-Yun (2006). A novel clustering method for examining and analyzing the intellectual structure of a scholarly field. Journal of the Korean Society for Information Management, 23(4), 215-231. https://doi.org/10.3743/kosim.2006.23.4.215

Lee, Won-Goo, Shin, Sung-Ho, Kim, Kwang-Young, Jeong, Do-Heon, Yoon, Hwa-Mook, Sung, Won-Kyung, & Lee, Min-Ho (2011). Semi-automatic management of classification scheme with interoperability. The Journal of the Korea Contents Association, 11(12), 466-474. https://doi.org/10.5392/jkca.2011.11.12.466