효과적인 학습 전략을 위한 컨텐츠 주제어 군집 방법

김준우 O*, 주지영**, 홍성용**, 이문용**, 윤완철**·
*한국기술교육대학교, **한국과학기술원
kimjunwoo@kaist.ac.kr, {jyjoo, gosyhong, munyi, wcyoon}@kaist.ac.kr

Contents Keyword Clustering for Effective Learning Strategy

Junwoo Kim O*, Jiyoung Joo**, Seongyong Hong**, Munyong Yi**, Wancheol Yoon**
*Korea University of Technology and Education, **Korea Advanced Institute of Science and Technology(KAIST)

요 약

e-Learning 환경에서 학습자들은 온라인 공간에 존재하는 다양한 학습 컨텐츠들을 이용하게 된다. 온라인 상의 컨텐츠들은 일반적으로 그 양이 방대하여 사용자가 체계적으로 이용하기 어렵기 때문에 최근에는 각 컨텐츠의 특성을 나타내기 위하여 키워드나 태그 등의 메타데이터를 함께 부여하는 경우가 많다. 본 논문에서는 학습 컨텐츠에 부여된 주제어들에 군집 분석을 적용하여 분류 체계를 생성하고, 이를 효과적인 학습 관리에 이용할 것을 제안한다. 또한, 학습 컨텐츠의 경우, 지나치게 많은 키워드에 연관된 것보다는 소수의 키워드에 연관된 것일수록 초기 학습에 유용하다는 가정 하에 기존의 분석 방법을 수정하는 방법을 고찰하였다.

1. 서 론

정보통신 기술의 발전과 함께 교육분야에서는 컴퓨터와 인터넷을 활용한 이러닝(e-Learning) 이 등장하였고, 이는 교수자와 학습자 간에 존재하는 시공간적인 장벽을 극복할 수 있는 방법으로 각광받고 있다[1]. 초창기 이러닝 컨텐츠와 관련 시스템들은 대부분 기존에 오프라인 교육 환경에서 사용되던 컨텐츠들을 단지 텍스트, 그림 또는 플래쉬 등의 형태로 가공하여 온라인 환경에 업로드 해 두는 'just-put-it-on-the-web' 방식의 접근에 그쳐왔다[1][2]. 하지만 최근에는 정보통신 기술을 최대한 활용하고, 오프라인 교육과는 차별화된 교육 환경을 제공하기 위하여 첨단 기술을 적용한 학습 컨텐츠를 소개하거나[3], 온라인 학습 환경에서 발생하는 다양한 데이터를 수집, 분석하여 활용하는 학습 전략을 제안하는 연구가 활발하다[1][4]. 특히 후자의 경우에는 여러 가지 데이터마이닝 기법을 적용하여 학습 효과 극대화를 위한 효과적인 학습 전략이나 학습 컨텐츠 추천 등을 목표로 하는 경우가 많다[1][5].

일반적으로 한 분야 또는 한 과목의 학습 내용은 세부 개념 또는 키워드들로 구성되고 초기 기본 학습을 위해서는 먼저, 단일 개념 또는 키워드에 대한 설명이나 정의, 때로는 예제 등이 학습자에게 제공된다. 하지만 Merrill 이 지적한 바와 같이, 학습 효과 극대화를 위해서는 실생활과 관련된 학습 컨텐츠나 문제 상황을 접하면서 여러 가지 개념들을 동시에 적용 및 고려하는 과정을 거치는 것이 필요하다[6]. 이러한 구성주의 패러다임에서는 개인의 자기주도형 학습경험과 여러 지식을 스스로 접목하면서 구성해 나가는 과정을 중시한다[3]. 이러한 맥락에서, 단일 개념 또는 키워드에 대한 기본 학습 외에 복수의 개념 또는 키워드와 연관된 컨텐츠들이 읽을 거리(article) 나 응용(application)문제 등의 형태로 제공되기도 한다.

학습 컨텐츠와 세부 개념들이 많은 경우에는 이러한 개념들의 연관성을 파악하기 어렵고, 학습자들이 학습 컨텐츠를 체계적으로 이용하기 곤란하다. 또한, 학습 관리자나 컨텐츠 저작자 입장에서도 현재 학습 환경의 특성과 향후 보완할 점 등을 파악하기 힘들다.

본 논문에서는 먼저, 어떤 학습 분야를 구성하는 세부 학습 주제 또는 키워드들의 연관 관계를 이용하여 생성한 분류 체계(taxonomy)를 학습 전략 수립에 사용하는 방안에 대해 고찰한 후, 이러한 분류 체계 생성 방법에 대해 논의해보고자 한다.

2. 관련 연구

본 논문에서는 복수의 개념 또는 키워드와 연관된 컨텐츠들을 분석하여 해당 분야의 지식 구성 체계를 분석하고, 세부 개념들 간의 연관성을 추출하고자 한다. 예를 들어, 최근 온라인 상의 방대한 컨텐츠들의 경우, 관련 키워드나 태그와 같은 메타 데이터를 부여하여 그 활용도를 높이는 경우가 많은데, 이렇게 연관 키워드 관련된 메타 데이터가 부여된 학습 컨텐츠들은 모두 이러한 방법으로 분석이 가능하다[7].

복수 개의 연관 개념 또는 키워드를 가진 컨텐츠의

본 연구는 산업원천기술개발사업(10035166: 창의적 인재육성을 위한 지능형 튜터링 시스템 개발)의 지원을 받아 수행되었음.

경우, 일종의 transaction 데이터로 볼 수 있다. 이 경우, 개념 또는 키워드 $\{i_1, i_2, \cdots, i_n\}$ 은 항목 (item)으로, 개별 컨텐츠 $\{T_1, T_2, \cdots, T_m\}$ 의 경우 transaction 을 형성하는 것으로 볼 수 있다. 이러한 transaction 데이터에서 추출되는 지식 또는 패턴 중 가장 대표적인 것은 연관(Association) 규칙이다. 연관 규칙은 일반적으로 (X → Y) 형태로 표현되며, 이 때 X, Y 는 각각 전체 항목 집합 (item set) $\{i_1, i_2, \cdots, i_n\}$ 의 부분 집합들이다. 규칙 (X → Y) 가 의미하는 바는 항목 집합 X를 갖는 transaction 의 경우, 항목 집합 Y 도 포함하는 경우가 많다는 것으로, 연관 규칙을 평가하는 척도로는 지지도(support), 신뢰도(confidence) 등이 사용된다[8][9]. 탐사된 연관 규칙은 어떤 항목을 선택한 고객에게 다른 항목을 추천하는 협업 필터링(collaborative filtering) 기법에 종종 사용되며, Hsu와 같이 학습자들에게 적절한 학습 컨텐츠를 추천하는 데에도 적용된 사례가 있다[5].

일반적으로 transaction 데이터에 대한 분석은 연관 규칙 탐사를 목적으로 하는 경우가 많지만, transaction 데이터에 군집(clustering) 분석을 적용하는 연구도 과거 수행되어 왔다. 군집 분석은 유사한 데이터들끼리 군집(cluster)을 이루도록 분류(grouping) 하는 것을 목표로 하며, 이 때 같은 군 내의 데이터들 간의 유사도(similarity)는 최대화하고, 다른 군 간의 데이터들 간의 유사도는 최소화되어야 한다[11]. 군집 분석은 마케팅, e-비즈니스 및 웹 마이닝 등에 광범위하게 적용되어 왔으며, 대표적으로는 K-means 등과 같은 프로토타입 기반(prototype-based) 군집 방법, 계층(hierarchical) 군집 방법, 밀도 기반(density-based) 군집 방법 등으로 분류된다[9]. 일반적으로 군집 분석은 여러 필드(field)가 하나의 레코드(record)를 묘사하는 record 형식의 데이터에 적용되는 경우가 많다.

트랜잭션 데이터는 한 개 레코드가 단지 여러 항목들의 집합을 의미하는 하나의 필드만을 갖는다는 점에서 일반 레코드 데이터와 구별되고, 여기에 군집 분석을 적용하는 목적은 크게 두 가지로 분류된다. 첫 번째는 항목들 중, 서로 연관성이 많다고 생각되는 것들끼리 군집하는 것이다. 이는 서로 유사한 레코드들끼리 군집하는 것을 목표로 하는 종래의 군집 분석 방법과는 다소 상이한 방법으로 볼 수 있다.

Han et al. [11] 은 트랜잭션 데이터에서 먼저 빈발 항목 집합(frequent item set) 을 탐사한 후, 이들을 하이퍼그래프(hypergraph) 형태로 나타내고, 그 안의 하이퍼에지(hyperedge) 들을 적절히 분할함으로써 함께 선택되는 경우가 많은 항목들끼리 군집화하는 방법을 제안하였다. 반면, 이후의 항목 군집 관련 연구들에서는 계층 군집 기법을 적용하여 항목들 간의 분류 체계까지 함께 생성하는 방법들이 제안되어 왔다. 계층 군집은 크게 병합 형(agglomerative)과 분할 형(Divisive) 으로 나뉘어지는데, 병합 형이 일반적으로 더 많이 사용되며, 이 경우, 최초에 모든 항목들은 개별 군집으로 취급되며, 현재 군집들 중, 가장 유사한 두 개를 골라 병합하여 하나의 군집으로 만드는 작업을 반복해나가게 된다. 따라서, 두 군집 간의 유사도를 정의하는 방법이 중요한 이슈가 되고, 일반적으로는 단일 링크, 완전 링크, 그룹 평균 등의 방법이 사용된다[9]. 계층 군집이 완료되면, 각 항목들간의 관계가 계통도(dendrogram) 형태로 표현되기 때문에, 항목들 간의 연관성을 체계적으로 알아볼 수 있다는 장점이 있다. 트랜잭션 데이터의 항목을 계층 군집하는 연구들은 군집 간 유사도 척도를 어떻게 정의하는가에 따라 조금씩 다른 특징을 가지고 있다.

Sarwar et al.,[10] 은 먼저 고객들이 각 항목에 대한 선호도 평가를 정량적으로 하여, 항목 하나에 대한 전체 고객들의 평가 점수를 벡터로 표현한 후, 이 벡터들 간의 코사인 유사도 등을 계산하여 항목 간의 유사도로 사용하였다. Yun et al.,[12] 은 항목들 간에 기존의 분류 체계(taxonomy)가 있는 경우, 분류 체계 상의 유사도와 트랜잭션 데이터 상의 유사도를 함께 고려하여 항목 간의 유사도를 계산할 것을 제안하였다. 본 논문과 같이 컨텐츠의 키워드들을 군집한 사례로는 최근의 Tsui et al. [13]을 찾아볼 수 있다. 이 연구에서는 먼저 50개의 최근 IT 관련 키워드를 선별한 후, 과거 10년 동안 IT 관련 잡지나 기사 등의 컨텐츠를 분석하여, 키워드들이 함께 등장한 횟수(Co-occurrence)를 키워드들간의 유사도로 사용하였다. 계층 군집 결과로는 그림 1과 같이 IT 키워드들간의 연관성을 보여주는 계통도가 생성되었다.

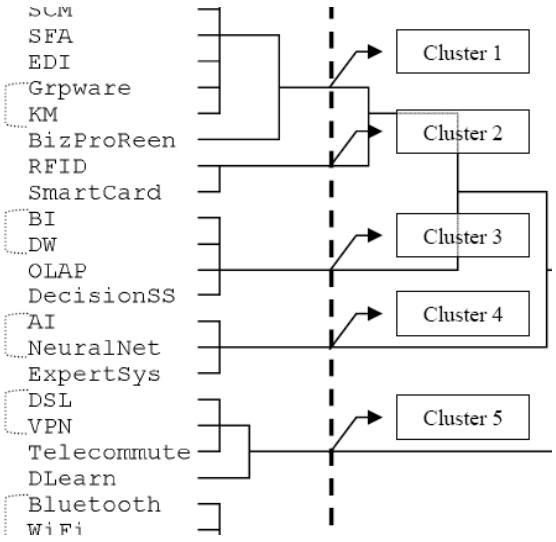

그림 1. IT 키워드 계층 군집 결과 (일부)

트랜잭션 데이터에 군집 분석을 적용하는 두 번째 목적은 일반적인 레코드 데이터 군집 분석에서와

마찬가지로 유사한 트랜잭션 들을 군집하는 것으로, 먼저 항목들을 군집한 후, 각 트랜잭션에 대하여 각 항목 군집과의 연관성 정도를 산출하여 군집하는 방법 [11], 큰 빈발 항목 집합들을 탐사하여 이들을 이용하여 트랜잭션들을 군집하는 방법 등이 있다 [14]. 본 논문에서는 첫 번째로 언급한 것과 같이 항목들을 계층 군집하여 학습 환경에 이용하는 방안에 대해 고찰하고자 한다.

3. 항목 분류 체계의 교육적 활용

그림 1의 계통도는 계층 군집을 통하여 여러 IT 관련 키워드들의 연관 관계를 보여주고 있다. 예를 들어, Cluster2에는 RFID, SmartCard 두 개의 키워드가 소속되어 있고, Cluster 3에는 BI(Business Intelligence), DW(Data warehouse), OLAP, DecisionSS(Decision Support System) 의 네 개 키워드가 소속되어 있다. 이 두 군에 대해 간단히 해석해 보면, Cluster2의 경우에는 최근 들어 각광을 받고 있는 유비쿼터스 컴퓨팅 환경에 관련된 군집이고, Cluster3은 기업에서 데이터를 수집, 가공 및 분석하여 실제 운영 관련 의사 결정에 사용하기 위한 기술 관련 군집이라는 것을 알 수 있다.

Tsui et al.에서는 컨텐츠의 연관 키워드들 간의 연관성 및 계통 구조 파악을 위하여 계층 군집을 적용하였다[13]. 좀 더 발전하여 이러한 분석은 보다 효과적인 학습 환경을 학습자에게 제공하는 데도 도움이 될 것으로 생각된다. 예를 들어 최근 IT 관련 동향에 대하여 웹 문서 검색을 통한 조사 및 학습을 하고자 하는 학습자가 있다고 가정하자. IT관련 용어들은 서로 다른 키워드들끼리 연관되어 있는 경우가 많고, 종종 다른 키워드에 대한 배경 지식이 있어야 이해가 원활한 경우도 많다. 따라서 최근 IT 관련 키워드들에 대해 어느 정도의 배경 지식 없이 무작정 관련 문서들을 검색하는 것은 원하는 목적을 달성하는데 효율적이지 못할 수 있다. 반면, 그림 1과 같은 키워드 간 연관 정보를 학습자에게 사전에 제공하고, 각 군집에 대한 간단한 설명을 제공하는 경우에는 학습자가 보다 효과적인 정보 탐색 및 학습을 하는 것이 가능할 것이다. 두 번째로, 그림 1의 DLearn (Distance Learning)키워드에 관심이 있는 학습자를 가정해보자. 이 키워드는 DSL(Digital Subscriber Line), VPN(Virtual Private Network), Telecommute (Telecommuting)와 함께 Cluster5에 소속되어 있는데, 계통도에 표현되어 있듯이, 이 네 키워드 중, DSL, VPN, Telecommute 간의 연관성이 상대적으로 크고, DLearn은 앞의 세 키워드에 부가적으로 연관되어 있음을 알 수 있다. 따라서 DLearn 키워드에 대한 학습을 원하는 학습자에게는 먼저 DSL, VPN, Telecommute 키워드들에 대한 정보를 배경 지식으로 학습한 다음, DLearn을 학습할 것을 추천할 수 있다.

그림 1의 계통도는 50개의 IT키워드를 항목으로, 그리고 과거 10년간 6개 IT관련지들의 기사 콘텐츠들을 트랜잭션으로 하여 생성한 키워드 계통도이다. 하지만 다른 형태의 컨텐츠에 적용하기에 따라 다양한 방법의 활용도 가능할 것으로 생각된다. 예컨대, 여러 학습 키워드나 학습 단원의 지식들을 동시에 요구하는 문제들이 출제되는 과목을 생각해보자. 각 문제를 트랜잭션으로, 그리고 학습 키워드나 단원을 항목으로 사용하여, 항목들간의 연관성을 계통도로 나타내면, 현재 이 과목에서는 다양한 학습 주제들이 주로 어떤 식으로 융합되어 응용되는지, 이 과목의 출제 경향은 어떠한지를 파악할 수 있어, 교수-학습에 도움이 될 수 있다. 또한 여러 논문이나 특허의 관련 키워드를 분석하여, 해당 분야의 지식이 어떤 구조를 가지고 있고, 유용한 문헌을 탐색하거나, 현재 취약한 연구 분야를 찾아내어 새로운 연구 주제를 선정하는 데도 도움이 될 수 있을 것이다.

4. 동시발생 키워드 기반 계층 그룹

Tsui et al.에서는 단순히 두 키워드들이 여러 컨텐츠에서 동시 발생(co-occurrence)하는 횟수를 두 키워드 간 유사도로 보았다[13]. 예를 들어, A, B, C, D, E의 5개 키워드와 표 1의 T_1, T_2, …, T_{10} 과 같은 컨텐츠가 있을 때, 키워드를 계층 군집하는 것을 생각해보자.

표 1. 컨텐츠-연관 키워드 목록 테이블

컨텐츠	연관 키워드	컨텐츠	연관 키워드
T_1	A, C	T_6	B, D, E
T_2	B, C	T_7	D, E
T_3	A, C, D	T_8	C, D
T_4	A, B, D, E	T_9	A, B, D, E
T_5	A, C, D, E	T_{10}	A, C, D, E

이를 위해 필요한 것은 먼저, 표 2와 같은 두 키워드 간 동시 등장 횟수를 행렬로 정리하는 것이다. 표 2에서 동시 등장하는 횟수가 가장 많은 키워드 쌍은 A와 D 임을 알 수 있다.

표 2. 동시발생 매트릭스 테이블 (예1)

	A	B	C	D	E
A	-	2	4	5	4
B	2	-	1	3	3
C	4	1	-	4	3
D	5	3	4	-	4
E	4	3	3	4	-

트랜잭션 데이터에서 항목을 군집화할 때는 기본적으로 함께 등장하는 횟수가 많은 항목들끼리 연관도가 큰 것으로 본다. Tsui et al.에서도 이 경우, 먼저 A, D를 병합하여, (A, D), (B), (C), (E) 의 네 개 군집이 생기게 된다[13]. 그 다음에는 다시 이 네 군간의 유사도를 측정하게 되는데, 두 개 이상의 항목이 있는 군과 다른 군의 유사도는 개별 항목들 간의 동시 등장 횟수 중, 두 군집에서 조합되는 모든 항목 쌍들의 동시 등장 횟수 평균을 이용하는 그룹 평균을 사용하였다. 이런 방법을 통하여, 한 번에 두 개 군집을 병합하면서 최종적으로는 한 개의 군집만 남을 때까지 병합을 반복하면 그림 1과 같은 키워드들의 계통도를 생성할 수 있다.

5. 개선된 항목 계층 군집 방법

앞에서 설명한 방법은 그 의미와 목적이 직관적이지만, 특정 성격을 가진 컨텐츠의 경우에는 소기의 목적과 다소 부합하지 않는 결과를 생성할 수 있다. 따라서 본 장에서는 학습 컨텐츠에 적용할 경우, 고려할 수 있는 사항들에 대하여 설명한다.

5.1 가중치를 포함한 유사도 측정

먼저, 표 2에서는 가장 빈발하게 같이 등장하는 A, D 키워드 쌍이 가장 유사하다고 판단하게 된다. 하지만, 표 1을 볼 경우, A, D 가 가장 먼저 병합되는 것에는 의문이 따를 수 있다.

즉, A, D의 경우 T_3 , T_4 , T_5 , T_9 , T_{10} 의 5개 컨텐츠에서 함께 연관되어 있지만, 이 컨텐츠들은 첫 번째 것은 연관 키워드가 총 3개이고, 나머지는 모두 4개씩의 키워드와 연관되어 있다. 학습 컨텐츠의 경우, 많은 키워드와 연관되어 있는 것은 난이도가 높은 컨텐츠이거나, 개별 키워드들에 대한 자세한 설명을 제공하기보다는 여러 키워드들을 개괄적으로 설명하여, 초기 학습에는 적절치 못할 수도 있다. 즉, A, D를 제일 먼저 연관시켜 이들 간의 학습을 추천하는 것은 효과적이지 못할 가능성이 커진다.

반면, A, C 쌍의 경우는 T_1 , T_3 , T_5 , T_{10} 의 4개 컨텐츠에서 함께 연관되어, A, D 보다는 동시에 나오는 빈도가 낮다. 하지만, T_1 은 A, D 키워드에만 연관되어 있고, 이러한 컨텐츠는 A, D 키워드와의 순수한 연관성이 상대적으로 높을 가능성이 크다.

학습자의 입장에서는 여러 키워드들이 한꺼번에 포함된 컨텐츠를 처음부터 이용하기보다는 맨 처음에는 적은 키워드에 대한 설명을 접하는 것이 유용할 수 있고, 이 경우에는 A, C와 같은 쌍이 초기에 병합되는 것이 유리하다. 이러한 점을 고려하여, 본 논문에서는 키워드 i_a , i_b 의 유사도를 단순히 동시 등장 빈도를 세어 측정하는 것 대신, 등장 빈도와 함께 연관의 순수도를 함께 고려하는 것을 제안하고, 먼저 아래와 같은 가중치를 고려한 동시 등장 횟수를 고안하였다.

$$\text{Weighted Co-occurrence}(i_a, i_b) = \sum_{k=1}^{m} \frac{2}{|T_k|} \quad \cdots\cdots(1)$$

$|T_k|$ 는 k 번째 컨텐츠의 총 연관 키워드 개수이고, 위 (1)은 어떤 두 개의 키워드가 동시에 등장하는 컨텐츠 1개가 있을 때, 두 키워드의 동시 등장 횟수를 1씩 증가시키지 말고, 두 키워드가 그 컨텐츠의 연관 키워드 목록에서 차지하는 비율만큼 증가시키게 된다. 따라서, 정확하게 그 두 키워드하고만 연관된 컨텐츠의 경우에는 횟수가 1씩 증가하지만, 연관된 키워드가 많아지면 횟수의 증가량은 작아진다. 앞의 (A, D)의 경우 (1)을 이용하여 측정한 횟수는 $\frac{2}{3}+\frac{1}{2}\times 4 = 2.66$, (A, C)는 $1+\frac{2}{3}+\frac{1}{2}\times 2 = 2.66$ 으로 두 쌍의 가중치를 고려한 등장 횟수는 동일해진다. 즉, 식 (1)은 많은 키워드들과 관련된 컨텐츠들 보다 해당 키워드들과 순수하게 연관된 컨텐츠에 가중치를 둔다. 하지만, (1)식만을 사용할 경우, 순수한 연관이 일어나는 비율은 작지만, 많은 컨텐츠들과 연관된 키워드 쌍의 유사도는 여전히 높게 나타날 가능성이 있다. 따라서, (1)의 값을 다시 해당 키워드 쌍의 지지도로 나눈 값을 유사도 척도로 사용하는 것을 제안한다.

$$\text{Sim}(i_a, i_b) = \text{식(1)} \;/\; \text{Sup}(i_a, i_b) \quad \cdots\cdots(2)$$

위 (2)를 사용한 키워드 간 유사도는 (A, D)의 경우 2.66/5 = 0.53, (A, C)의 경우 0.67로 나타나, 이제 최초 병합 시, (A, D)보다는 (A, C)의 병합을 선호하게 되고, 학습자 입장에서 학습 경로를 선정하기 위한 학습 컨텐츠 연관 키워드 구조 파악에 보다 적합해진다.

5.2 동시발생 유사도 그룹

두 키워드 간의 유사도가 정의된 후에는 이를 반복적으로 사용하여 계통도를 만들어나가게 된다. 이 과정에서 개별 키워드 간의 비교가 아닌, 2개 이상의 키워드를 포함한 군집 간의 유사도를 정의하는 방법이

필요하다. 널리 쓰이는 단일 링크, 그룹 평균, 완전 링크에 대해 논의하기 위하여 A, B, …, F 의 6개 키워드가 표 3의 행렬과 같은 동시발생 매트릭스를 갖는다고 하자.

표 3. 동시발생 매트릭스 테이블 (예2)

	A	B	C	D	E	F
A	-	7	1	2	3	3
B	7	-	2	5	2	3
C	1	2	-	6	1	1
D	2	5	6	-	1	1
E	3	2	1	1	-	8
F	3	3	1	1	8	-

먼저, (A, B), (C, D), (E, F) 는 각각 동시 발생 횟수가 많은 키워드 쌍이므로, 계층 군집에서 초기에 병합될 것이다. 따라서 초기에 위의 세 군집이 발생한다. 그 다음에는 이 세 군집들간의 병합에 대해 고려하게 되는데, 먼저 표 3에서 (C, D), (E, F) 군집 간의 키워드들끼리는 동시 발생 횟수가 모두 1로 낮으므로, 두 군집은 상대적으로 나중에 병합될 것이다. 대신 이번에 수행할 절차는 (A, B)에 (C, D) 또는 (E, F)를 병합하는 것이 될 것이다.

여기서, (A, B)-(C, D)간의 유사도와 (A, B)-(E, F)간의 유사도에 대해 생각해보자. 먼저, (A, B), (C, D) 간의 유사도는 단일 링크로 측정할 경우, 양쪽 군집의 키워드 중, 가장 빈번하게 함께 등장하는 B, D 간 횟수인 5가 될 것이다. 그룹 평균을 사용한다면, A-C, A-D, B-C, B-D 간 횟수들의 평균인 2.5가 된다. 완전 링크를 사용하는 경우에는 양쪽 군집에서 가장 유사도가 적은 A-C 간 횟수인 1이 두 군집 간 유사도가 된다. 마찬가지 방법으로 (A, B)-(E, F) 간의 유사도도 세 가지 방법으로 계산한 값들을 표 4에서 비교해볼 수 있다.

표 4. (A, B) 군집과의 유사도

방법	(C, D)	(E, F)
단일 링크	5	3
그룹 평균	2.5	2.75
완전 링크	1	2

표 4에서 단일 링크를 사용할 때만, (C, D)를 먼저 (A, B)와 병합하게 되고, 그룹 평균이나 완전 링크를 사용할 때는 (E, F)의 유사도가 더 크므로 이 군집을 (A, B)와 병합하게 됨을 알 수 있다. 이러한 결과가 나오는 이유는 (C, D)의 경우에는 D 키워드만이 (A, B) 군집 내의 B와 연관성이 크고, 나머지는 키워드들 간의 연관성은 그리 크지 않기 때문이다. 반면, (E, F)의 경우에는 B-D 와 같이 연관성이 큰 쌍은 존재하지 않지만, 전체적으로는 연관성이 극히 작은 쌍도 없다.

즉, 단일 링크는 유사도가 큰 항목 쌍이 있는 경우, 나머지들과 관계없이 이에 영향을 많이 받는 특성이 있고, 전통적인 계층 군집 연구에서는 이 점이 올바르지 못한 군집을 만들 수 있다고 지적되지만, transaction 데이터의 항목이 학습 컨텐츠와 연관된 키워드들일 경우, 상황이 달라질 수 있다.

예를 들어, (A, B) 키워드와 관련된 학습을 끝낸 학습자의 경우, 먼저 B 와 연관성이 큰 D를 추천하면, 학습의 시너지 효과가 가장 클 것이다. 그리고, D에 대한 학습이 끝난 경우에는 D 와 같은 군집에 있는 C로 자연스럽게 학습 경로를 조정해가는 것이 가능하다. 반면, (A, B) 에 대한 학습을 끝낸 학습자에게 E나 F를 추천하는 것은 D보다 상대적으로 학습효과가 작아진다. 따라서, 학습 컨텐츠의 경우에는 (A, B)와 (C, D)를 병합하는 것이 바람직할 수 있다.

즉, 계층 군집의 목적에 따라, 군집 간의 병합 시, 전체적인 유사도 보다는 연관성이 큰 특정 항목들 간의 유사도만 고려하는 것이 필요할 수 있고, 이러한 경우에는 군집 간 유사도 척도로 단일 링크를 사용하는 것이 바람직할 것이다.

6. 결론 및 추후 과제

본 논문에서는 학습 컨텐츠에 부여된 키워드들에 계층 군집 방법을 적용하여 전체적인 계통도를 형성하고, 이를 이러닝 분야에 활용하기 위한 기초적인 방법에 대해 논의하고 있다.

키워드가 부여된 컨텐츠들은 각각 일종의 트랜잭션 데이터로 볼 수 있고, 기존에 연구되었던 분석 방법을 적용할 수도 있지만, 학습 컨텐츠 키워드의 경우에는 일반적인 데이터와는 조금 다른 특성이 존재한다. 특히 본 논문에서는 1)연관 키워드 개수가 많은 컨텐츠의 경우, 그 키워드들 간의 연관 관계가 적어지고, 난이도 등의 이유로 초기 병합에 크게 기여할 수 없도록 하는 것이 바람직함과, 2)학습의 시너지 효과를 위해서는 계층 군집 시, 단일 링크를 사용할 것을 제안하였다. 학습자가 초기에 단일 개념에 대해 학습하고, 추후에 개념들간의 연관성에 대해 학습하는 경우에 이러한 방법으로 생성한 계통도는 효과적으로 여러 키워드들을 학습해나갈 수 있는 학습 경로를 추천하는데 유용할 것으로 생각된다.

추후 연구 과제로는 실제 데이터에 이러한 방법을 적용하고, 생성된 계통도를 활용한 구체적인 학습 전략 수립 방안을 수립해보는 것을 연구하고 있다. 또한, 컨텐츠의 연관 키워드외 부가적인 정보, 예를 들어 컨텐츠의 난이도, 인기도 등을 함께 고려하는 분석 방법을 연구 중에 있다.

참고문헌

[1] Romero, C. and Ventura, S., "Educational data mining: A survey from 1995 to 2005", Expert Systems with Applications, Vol.33, No.1, pp.135-146, 2007.

[2] Brusilovsky, P. and Peylo, C., "Adaptive and Intelligent Web-based Educational Systems", International Journal of Artificial Intelligence in Education, Vol.13, pp.159-172, 2003.

[3] 김용훈, 이수웅, 이준석, 노경희, "혼합현실기반 이러닝 기술동향", 전자통신동향분석, 제24권, 제1호, pp.90-100, 2009.

[4] D'Mello, S. and Graesser, A., "Mining Bodily Patterns of Affective Experience during Learning", Proceedings of the 3rd International Conference on Educational Data Mining, pp.31-40, 2010.

[5] Hsu, M.-H., "A Personalized English Learning Recommender System for ESL Students", Expert Systems with Applications, Vol.34, No.1, pp.683-688, 2008.

[6] Merill, M.D., "First Principles of Instruction", Educational Technology Research and Development, Vol.50, No.3, pp.43-59, 2002.

[7] Deitel, P. and Deitel, H., "Internet & World Wide Web: How to Program, 4th e/d", Prentice Hall Press, 2007.

[8] Agrawal, R. and Srikant, R., "Fast Algorithms for Mining Association Rules in Large Databases", Proceedings of the 20th International Conference on Verery Large Data Bases, pp.478-499, 1994.

[9] Tan, P.-N., Steinbach, M. and Kumar, V., "Introduction to Data Mining", Addison Wesley, 2005.

[10] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J., "Item-Based Collaborative Filtering Recommendation Algorithms", Proceedings of the 10th International Conference on World Wide Web, pp.285-295, 2001.

[11] Han, E.-H., Karypis, G., Kumar, V. and Bamshad, M., "Clustering Based on Association Rule Hypergraphs", Proceedings of SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 1997.

[12] Yun, C.-H., Chuang, K.-T. and Chen, M.-S., "Clustering Item Data Sets with Association-Taxonomy Similarity", Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03), pp.697-700, 2003.

[13] Tsui, C.-J., Wang, P., Fleischmann, K.R., Sayeed, A.B. and Weinberg, A., "Building an IT Taxonomy with Co-occurrence Analysis, Hierarchical Clustering and Multidimensional Scaling", Proceedings of iConference, pp.247-256, 2010.

[14] Wang, K., Xu, C. and Liu, B., "Clustering transactions using large items", Proceedings of the 8th International Conference on Information and Knowledge Management, pp.483-490, 1999.