전자상거래에서의 제품의 구매 연관과 내용 유사도간의 상관관계 연구*

이 홍 주**

A Study of the Relationship between Purchase Association and Contents Similarity of Products in e-Commerce*

Hong Joo Lee**

■ Abstract ■

Many online stores provide relevant products in product pages or other pages to attract customer interests. Association rules based on customer purchases and personalized recommendations are most prominently used ways of providing relevant products. Though there have been many studies to apply tags of products as metadata of the products, there are few studies to investigate contents similarities between the products and the suggested products. Thus, this study collects books in purchase associations and their tags in Amazon.com and assesses the similarities between the books. We found out that the contents similarities based on tags are similar among business, literature, and computer networks. And the similarity is also similar among the relevant books that have different ranks.

Keyword : Association Rules, Recommendation, Contents Similarity, Electronic Commerce

논문투고일 : 2011년 03월 21일 논문수정완료일 : 2011년 04월 26일 논문게재확정일 : 2011년 05월 23일
* 이 논문은 2008년도 정부재원(교육과학기술부 인문사회연구역량강화사업비)으로 한국연구재단의 지원(NRF-2008-332-B00097)과 2009년도 가톨릭대학교 교비연구비의 지원으로 이루어졌음.
** 가톨릭대학교 경영학부

1. 서 론

전자상거래의 발달과 함께 한 온라인 상점에서 취급하는 제품의 수와 제품 카테고리는 크게 증가하였다[9]. 온라인 상점들은 고객들이 원하는 상품을 찾아가게 하기 위한 다양한 기능을 제공하고 있는데, 고객이 제품을 찾는 주요한 방식은 검색과 카테고리 브라우징이다[14]. 또한, 다양한 영역에 사용자가 관심 있어 할만한 상품을 노출시켜 고객들의 관심을 끌기위해 노력하고 있다. 홈페이지에 많은 상품을 노출시키거나, 상품상세 페이지에서도 많은 연관 상품을 제시하고 있다. 개인화 추천을 활용하여 고객이 관심 있어 할만한 상품을 따로 모아 보여주기도 한다. 연관 상품 제시를 통해 고객들의 관심을 끈다면, 이를 통해 유발된 관심이 고객의 고려와 구매의사결정에 영향을 미치게 된다[14, 26]. 연관 상품 제시에서 가장 많이 활용되는 방안은 고객들의 구매 데이터를 활용하여 함께 구매하는 경향이 있는 제품을 찾아서 보여주는 것이다. 온라인 상점에서 '이 제품을 구입한 분들은 다음 제품도 구입하셨습니다'와 같은 문구와 함께 구매연관 제품을 제품상세 페이지에서 보여주고 있다. 같은 분야의 새로운 제품이나 베스트셀러 등을 보여주기도 하며, 마케팅 목적으로 마케터에 의해 정해진 상품들이 노출되기도 한다.

이러한 연관 상품 제시는 목적지향적인 고객들과 그렇지 않은 고객들 모두에게 더 많은 상품의 방문과 더 오랜 사이트 체류를 유도하는 효과를 갖는다[14]. 간단한 연관 규칙[17]을 활용하는 구매연관 상품 제시에서 부터 복잡한 연산을 필요로하는 개인화 추천까지 다양한 방식의 알고리즘에 대한 연구들은 많이 이루어졌지만, 구매 연관관계를 통해 파악된 제품들 간의 내용 유사도나 개인화 추천을 통해 제시된 제품들 간의 내용 유사도에 대한 조사는 이루어지지 않았다. 구매 연관관계나 개인화 추천이 고객들의 구매행위와 선호도 표현 행위에 의해 유사한 상품을 찾기 때문에 제시된 상품들이 내용적으로 유사할 것으로 생각된다. 하지만, 구매연관 및 추천 관계에 놓인 상품들간의 내용 유사도에 대한 연구가 거의 이루어지지 않았으며, 구매 연관 상품 제시와 개인화 추천 상품 제시 방안을 연구하는데 시사점을 줄 수 있을 것이다.

따라서 본 연구에서는 구매 연관관계를 통해 연관 상품으로 제시된 상품들의 내용 유사도와 개인화 추천된 상품의 내용 유사도를 조사하였다. 이를 위해 Amazon.com에서 비즈니스, 문학, 컴퓨터 네트워크 분야의 베스트셀러와 이에 대한 구매 연관 상품을 수집하였으며, 각 도서에 대해 사용자들이 입력한 태그 또한 수집하였다. 내용 유사도를 계산하기 위하여 연관 도서간의 태그 일치도를 계산하여 활용하였다. 각 분야별로 연관 상품의 유사도와 연관 상품 순서에 따른 유사도를 비교하였다. 또한 개인화 추천된 도서간의 내용 유사도도 동일한 방안으로 조사하였다. 이를 통해 구매 연관관계에 있는 상품의 내용 유사도와 개인화 추천관계에 있는 상품의 내용 유사도를 파악할 수 있으며, 연관 상품 제시 알고리즘과 고객 행동의 관계에 대한 이해를 높일 수 있을 것이다.

제 2장은 연관 상품 제시와 태그의 활용에 대한 관련 연구를 정리하였으며, 제 3장은 본 연구에 활용된 자료의 수집방안과 수집된 자료를 정리하였다. 제 4장에서 수집된 자료를 활용하여 분석한 결과를 제시하였으며, 제 5장에서 분석 결과에 대한 토의와 본 연구의 결론을 제시하였다.

2. 문헌 연구

2.1 연관 상품 제시

온라인 상점이나 정보제공업체에서는 고객들에게 다양한 방식으로 연관 상품을 제공하고 있다. 연관 상품을 제공하는 이유는 고객들이 관심 상품을 손쉽게 찾을 수 있도록 하는 목적과 노출된 상품에 대한 관심유발을 통해 고객의 구매를 유도하려는 데 있다. <표 1>에 국내외 대표적인 온라인 업체들의 연관 상품 제시 방안을 정리하였다. 고객의 구매나 방문 연관관계를 활용하여 연관상품을 제시하는 경우가 대표적이며, 추가적으로 고객

〈표 1〉 대표적 온라인 상점의 연관 상품 제시 방안

온라인 상점	연관 상품 제시 방안
Amazon.com	Frequently Buy Together(자주 함께 구매되는 상품들) Customers who bought this item also bought(이 상품을 구매한 고객들이 구매한 다른 상품) Customers also bought items by(고객들이 구매한 다른 저자들의 상품) Customers who viewed this item also viewed(이 상품을 방문한 고객들이 방문한 다른 상품) What do customers ultimately buy after viewing this item(이 상품을 본 고객들이 실제로 구매한 상품) Listmania!(이 상품이 포함된 상품 리스트) So You'd like to … (당신 좋아할 만한 상품) Customers who bought items in your recent history also bought(최근 방문한 상품들을 구매한 고객들이 구매한 다른 상품) Today's recommendation for you(오늘의 추천 상품, 개인화) Recommended for you(추천 상품, 개인화)
Netflix	More like this item(이 영화와 비슷한 다른 영화) Similar titles available to watch(검색시 제공, 유사 영화) Top 10 for the user(추천 영화 Top 10, 개인화) Suggestions for you(추천 영화, 개인화)
알라딘 Aladdin.co.kr	이 책을 구입한 분들은 다음 책도 구입하셨습니다(구매연관 상품) 관련 상품(DVD, 원서/번역서) 오늘 본 상품과 관련 추천 상품 추천 마법사-마법사의 선택(상품 추천) 추천 마법사-서재이웃의 선택(유사성향 사용자를 고려한 상품추천)
GS e-shop	브랜드 추천 베스트(방문 상품 브랜드의 베스트 상품 추천) Best of Best(방문 상품 카테고리의 베스트셀러 상품 추천)

의 구매 내역이나 선호도에 기반을 두어 개인화된 추천을 제공하는 것과 고객들이 작성한 상품 리스트를 제공하는 것이 활용되고 있다. 또한 방문한 상품의 분야별 베스트셀러나 방문한 상품 브랜드의 베스트셀러 상품을 제공하기도 한다. 이외에는 온라인 상점이 마케팅 중인 제품을 상품페이지에 노출시킬 수 있으며, 오픈 마켓인 경우에는 상품을 판매하는 업체가 노출 되는 연관 상품을 정할 수도 있다.

이홍주(2008)는 온라인 서점의 클릭스트림 데이터 분석을 통해 상품상세 페이지에서 연관 상품을 노출 시킬 경우 고객이 연관 상품을 클릭하는 비율이 높아지며, 이렇게 방문한 제품에 대한 고려와 구매로 연결되는 경향이 있는 것을 확인하였다[14].

많은 연구들이 상품의 구매 연관관계를 온라인 상점에서 연관제품 제시에 활용하기 위한 방안을 제시하였다. 대표적으로는 구매 연관관계를 개인화 추천 기법과 함께 적용하였다.

김재경 외(2007)은 전체 고객의 구매 연관규칙과 유사선호고객의 상품그룹을 혼합하여 개인화 추천을 수행하는 FREPIRS 시스템을 제안하였다[4]. 김지혜, 박두순(2006) 또한 개선된 Apriori 알고리즘과 아이템 대 아이템 상관관계를 고려하여 개인화 추천을 수행하는 알고리즘을 제시하였다[5]. 안현철 외(2006)은 고객의 인구통계 및 구매 정보에 연관관계 기법과 분류기법을 적용하는 추천 시스템의 모형을 제시하였으며, 온라인 설문을 통해 제시한 기법의 유용성을 검증하였다[10].

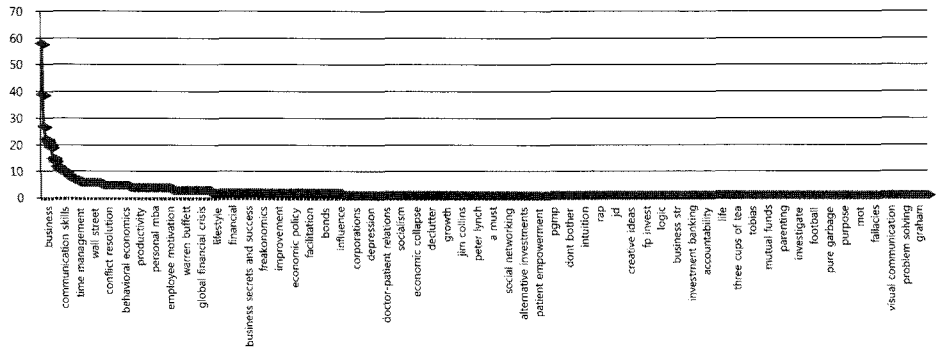
개인화 추천이외에도 상품의 가치 측정과 상품 검색에 연관관계를 활용하려는 연구들이 있었다. 황인수(2004)는 연관관계 규칙을 활용하여 상품 네트워크를 만들었으며, 교차판매 등을 고려한 상품의 수익예측 알고리즘을 제시하였다[15].

컴퓨터 시뮬레이션을 통해 제시한 알고리즘이 상품 가치를 독립적으로 계산하는 것보다 유용하다는 것을 보였다. 황현숙, 어윤양(2002)는 연관 마이닝과 고객 선호도에 기반을 두어 고객이 입력한 키워드 및 속성에 대한 가중치를 가지고 상품을 검색하는 시스템을 제안하였다[16].

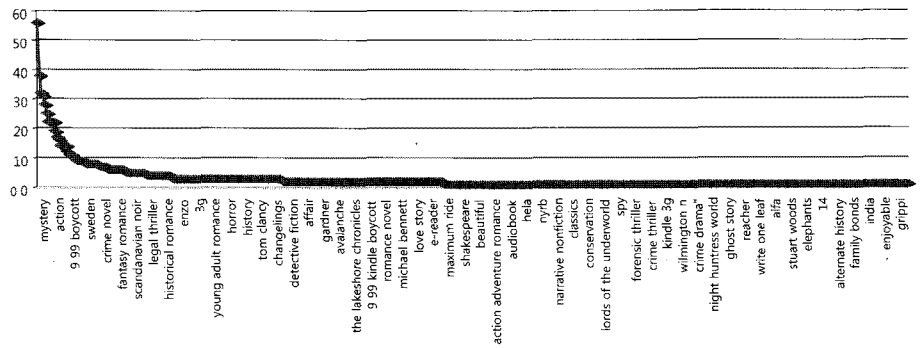
2.2 태그의 활용

사용자들의 참여를 통해 더욱 좋은 서비스를 제공할 수 있게 되는 것이 웹 2.0의 주요한 특징 중 하나이다. 많은 사용자들이 정보를 등록하고 태그를 등록할수록 각 정보에 대한 유용한 메타데이터가 등록될 가능성이 커지며, 이를 통해 연관 있는

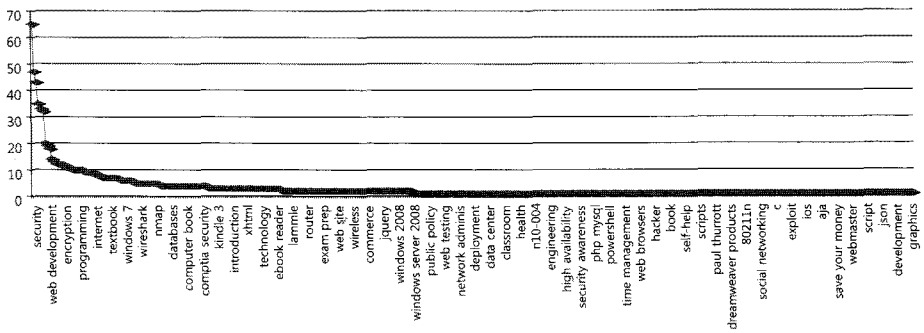
다른 정보들을 매개할 수 있게 된다. 다양한 사람들이 참여하여 정보를 공유하거나 생성하기 때문에 이미 분류된 기준(taxonomy)를 따르기 보다는 각 정보에 대한 메타데이터를 사용자들이 직접 입력하여 분류하는 방식(folksonomy)이 적합하였고 이를 위해 태그를 활용하였다[22]. 웹 블로그, 커뮤니티 글, 동영상, 사진, 음악 등 다양한 정보들에 작성자가 직접 입력한 태그들을 쉽게 접할 수 있으며, 입력된 태그를 선택하면 동일한 태그를 포함한 다른 웹페이지나 정보들을 손쉽게 파악할 수 있게 된다[2]. 메타 데이터로 입력되는 태그들을 통해 해당하는 정보가 속하는 분야를 만들어 가는



[그림 1] 비즈니스 분야 태그 분포도



[그림 2] 문학 분야 태그 분포도



[그림 3] 컴퓨터 네트워크 분야 태그 분포도

집단구조는 기존의 관리자에 의해서 정의되는 분류구조가 가지고 있는 정적인 분류 문제를 해결할 수 있으며, 동일한 태그나 유사한 태그를 가진 정보를 찾아가는 통로역할을 제공한다[22].

일반적으로 컨텐츠 태깅에 활용되는 태그를 빈도수로 정렬하면 본 연구에 활용된 태그 빈도수 [그림 1]~[그림 3]처럼 Power law 곡선 형태를 띤다[18, 19]. Golder and Huberman(2006)은 협업 태깅 사이트에서 활용되는 태그를 정보내용, 정보유형, 정보작성자, 정보분류, 정보품질, 자기정보, 관련업무로 분류하였다[18]. 대부분의 태그는 정보내용에 해당하며 정보가 담고 있는 내용을 기술한다. 정보유형은 정보가 블로그인지 동영상인지와 같은 유형에 대한 것이며, 정보 분류는 정보가 속한 분야에 대한 정보이다. 정보품질은 정보의 품질이나 특징에 대한 것으로 '재미있는'이나 '바보같은' 등이 그 예이다.

태그를 다양한 영역에 활용하려는 연구들이 있어왔다. 크게 구분하면 컨텐츠 검색과 컨텐츠 추천에 태그가 많이 활용되어 왔다. 강상욱 외(2010)과 Penev and Wong(2008)은 웹페이지의 태그를 활용하여 유사한 웹페지를 찾는 방안을 고안하였다. 김은희, 정영미(2010)은 블로그 검색을 위해 블로그의 본문 이외에 사용자 태그를 활용하였을 때 검색 성능이 향상되는 것을 보였다[3]. 박수진 외(2010)은 소셜 북마크에서 웹 페이지 공유자 수뿐만이 아니고 활용된 태그들의 가중치를 활용하여 검색결과를 제공하는 것이 효율적인 검색 결과를 도출하는데 도움이 되는 것을 보였다[8]. 엄태영 외(2010)은 소셜북마크의 검색결과를 개인화하여 재순위화하기 위하여 사용자가 입력한 태그들과 유사한 다른 태그들을 활용하였다[11].

많은 연구들이 유사한 웹 페이지나 정보를 찾아 추천하는데 사용자가 입력한 태그를 활용하였다. 김형도(2009)는 태그를 활용하여 웹페이지를 추천하는 잠재 요인 모델에 기반을 둔 알고리즘을 제시하였다[7]. Rhie et al.(2010)은 공유된 동영상의 태그를 활용하여 협업 필터링에 기반을 둔 추천방안을 제시하였다[24]. 연철 외(2008)은 사용자 태그를 가지고 협업 필터링과 분류기법을 적용하여 소셜 북마크의 웹페이지를 추천하였으며, 이 기법은 협업 필터링이 가지고 있는 초기 사용자 문제를 보완해 줄 수 있는 것으로 나타났다[12]. Jiao and Cao (2007)은 태그를 하이브리드 필터링 방식에 적용하여 추천을 수행하는 시스템을 제시하였다[20].

김현우 외(2010)은 사용자들이 본문을 작성한 후에 태그를 입력하는 것이 번거로워 잘 입력하지 않기 때문에, 사용자들에게 적합한 태그를 추천하여 손쉽게 사용자 태그를 입력할 수 있게 도와주는 알고리즘에 대해 조사하였다[6]. Sigurbjörnsson and Zwol(2008)도 Flickr에서 다른 사용자들이 사진에 입력한 태그를 활용하여 사용자가 입력할 태그를 추천하는 방안을 제시하였다[27].

즉, 태그가 관련된 컨텐츠의 메타 데이터로서 활용될 수 있으며 컨텐츠의 내용에 대해서 대표하는 역할을 수행하고 있음을 확인할 수 있다.

3. 자료 수집

온라인 상점인 Amazon.com에서 분석에 필요한 자료를 수집하였다. 자료를 수집한 시점은 2011년 2월 2일이며, 비즈니스(Business and Inventing), 문학(Literature), 컴퓨터 네트워크 분야 베스트셀러 리스트에 올라있는 도서중에서 상위 30개 도서를 선정하였다. 위의 세 분야를 선택한 이유는 컴퓨터 네트워크 분야는 비즈니스와 문학 분야에 비해 내용범위가 좁기 때문에 선택하였으며, 문학분야는 상당히 다양한 영역을 다루는 큰 범위의 분야이기에 선택하였다. 비즈니스 분야는 보다는 내용범위가 작고 컴퓨터 네트워크 보다는 범위가 큰 것으로 생각해 볼 수 있기 때문에 선택하였다. 하드커버, 페이퍼백, eBook과 같이 도서 내용은 똑같으나 유형만 다른 경우는 제거하고 고유한 도서로 30개를 선정하였다.

<표 2>는 수집된 분야의 도서목록 중 Top 10베스트셀러 목록이다. 수집한 데이터는 베스트셀러 30개 도서 목록과 이 도서를 설명하기 위해 사용자들이 입력한 태그들 그리고 각 도서와 함께 구

〈표 2〉 Top 10 베스트셀러 리스트

비즈니스 분야	문학 분야	컴퓨터 네트워크 분야
The Investment Answer	Alone	CISSP All-in-One Exam Guide
The Big Short	Switched	JavaScript : The Definitive Guide
Strengths Finder 2.0	Water for Elephants	The Web Collection Revealed
The 4-Hour Workweek	The Girl Who Kicked the Hornet's Nest	CompTIA Security+
As One	Tick Tock	Group Policy
Outliers	Saving Rachel	How to Disappear
The 7 Habits of Highly Effective People	The Girl with the Dragon Tattoo	Nmap Network Scanning
Scorecasting	The Girl Who Played with Fire	Wireshark Network Analysis
A Guide to the Project Management Body of Knowledge	Lethal People	Creating a Web Site
Good to Great	Cutting for Stone	The Web Application Hacker's Handbook

〈표 3〉 분야별 수집 태그 정보

	비즈니스 분야	문학 분야	컴퓨터 네트워크
총 도서 수	180	180	180
총 고유 도서 수	159	158	166
총 태그 수	1430	1427	1252
총 고유 태그 수	638	501	409
도서별 평균 태그 개수	8.99	9.03	7.54
최다 활용된 태그	Business (58회)	Mystery (56회)	Security (65회)

매된 도서(Customers who bought this book also bought)로 제시된 5권의 책들과 이들의 태그를 수집하였다. 베스트셀러 도서의 상세 페이지 URL을 입력하면 이 책의 태그들과 구매연관 도서의 상세 페이지 URL과 태그를 수집하는 웹 크롤러를 활용하여 정보를 동시에 수집하였다.

도서 정보페이지에는 연관 태그 중에서 사용자들이 가장 많이 입력한 상위 10개의 태그가 제시되기 때문에 도서별로 최대 10개의 태그를 수집할 수 있었다. 베스트셀러 별로 5권의 구매연관 도서를 구하였으나 이 중에는 중복되는 도서들이 있어서 각 분야별 도서의 총 고유 도서 수는 180권보다 적었다.

또한, 추천 도서와의 비교를 위해 Amazon.com에서 Recommended books페이지를 통해 저자에게 추천된 30권의 책과 추천에 근간이 된 책의 태그 정보를 수집하였다. Amazon.com에서는 추천도서를 제시하면서 "____을 구매하였기 때문에 추천함" 혹은 "____을 선호하였기 때문에 추천함"과 같이 이유를 제시하고 있다. 따라서 추천된 도서 30권과 이의 근간이 된 도서들의 태그를 수집하여 두 도서간의 태그 유사도 또한 비교하였다.

저자에게 추천된 도서들이기 때문에 비즈니스와 컴퓨터 관련 도서들이었다. Amazon.com의 도서 추천은 구매 연관과 다르게 아이템 기반의 협업 필터링 방식을 활용하고 있다[21]. 아이템 기반의 협업 필터링은 사용자 기반의 협업 필터링과 다르게

〈표 4〉 추천도서 리스트 예시

추천 도서	추천 기준 도서
Predictably Irrational	The Black Swan
How to win every argument	Blink
Drive : The Surprising Truth about What Motivates Us	The Culture of Collaboration
Analytics at Work	Competing on Analytics
Resonate : Present Visual Stories that Transform Audience	Presentation Zen
The Elements of Statistical Learning	Data Mining
Visual Meetings	Testing 1-2-3
Here Comes Everybody	Enterprise 2.0
Long Tail	The Economics of Information Technology
Adventures of an IT Leader	Real Business of IT

아이템을 기준으로 유사한 아이템을 찾아 추천한다[25]. 유사한 아이템을 선정하는 기준은 사용자들의 구매여부나 선호도를 활용한다. 자료를 수집한 추천도서 리스트 예시와 수집된 태그 정보가 <표 4>와 <표 5>에 정리되어 있다.

〈표 5〉 추천도서 수집 태그 정보

	추천 도서
총 도서 수	60
총 고유 도서 수	47
총 태그 수	424
총 고유 태그 수	254
도서별 평균 태그 개수	9.021
최다 활용된 태그	Statistics, Data mining, Business(9회)

4. 분석 결과

구매 연관도서간의 태그를 기반으로 한 내용 유사도를 계산하기 위하여 아래 유사도 계산 방안들을 활용하였다. Dice 유사도(S_d)는 N_r를 연관도서의 태그 수, N_b를 기준도서의 태그 수, N_s를 두 도서의 공통 태그 수라고 하면 아래와 같이 계산된다.

$$S_d = \frac{2 \times N_s}{N_b + N_r}$$

Jaccard 유사도(Sj)는 N0r를 연관도서에만 존재하는 태그의 수, Nb0를 기준 도서에만 존재하는 태그의 수, Nbr를 두 도서에 공통으로 존재하는 태그의 수라고 하면 아래와 같이 계산된다.

$$S_j = \frac{N_{br}}{N_{0r} + N_{b0} + N_{br}}$$

Cosine 유사도(Sc)는 N_b를 기준도서의 태그 벡터, N_r을 연관도서의 태그 벡터로 정의하면 아래와 같이 계산된다.

$$S_c = \frac{N_b \cdot N_r}{\| N_b \| \cdot \| N_r \|} = \frac{\sum_{i=1}^{n} N_{bi} \times N_{ri}}{\sqrt{\sum_{i=1}^{n} N_{bi}^2} \times \sqrt{\sum_{i=1}^{n} N_{ri}^2}}$$

도서 분야별과 연관도서별 유사도를 정리한 게 <표 6>, <표 7>, <표 8>이며, <표 6>은 Dice 유사도를 <표 7>은 Jaccard 유사도를 <표 8>은 Cosine 유사도 값이다.

Dice와 Jaccard 유사도는 연관도서 전체의 유사도를 계산하였을 때, 비즈니스 분야가 제일 높고, 문학 분야 유사도가 제일 낮았다. Cosine 유사도의 경우에는 컴퓨터 네트워크가 제일 높았으며, 문학 분야가 제일 낮았다. 연관순위에 따른 유사도는 대체로 연관순위가 높을수록 유사도가 높았으며, 연관순위가 낮아질수록 유사도 값이 하락하는 것으로 볼 수 있다.

〈표 6〉 Dice 유사도

	연관 도서 1	연관 도서 2	연관 도서 3	연관 도서 4	연관 도서 5	연관 도서 전체
비즈니스	0.378	0.341	0.279	0.246	0.243	0.297
문학	0.315	0.239	0.222	0.219	0.161	0.231
네트워크	0.334	0.274	0.238	0.202	0.286	0.267

〈표 7〉 Jaccard 유사도

	연관 도서 1	연관 도서2	연관 도서 3	연관 도서 4	연관 도서 5	연관 도서 전체
비즈니스	0.249	0.227	0.187	0.158	0.155	0.195
문학	0.208	0.152	0.129	0.134	0.094	0.144
네트워크	0.220	0.176	0.160	0.124	0.196	0.175

〈표 8〉 Cosine유사도

	연관 도서 1	연관 도서2	연관 도서 3	연관 도서 4	연관 도서 5	연관 도서 전체
비즈니스	0.380	0.345	0.282	0.249	0.247	0.301
문학	0.374	0.290	0.244	0.266	0.184	0.271
네트워크	0.384	0.329	0.282	0.246	0.344	0.317

도서 분야별로 연관도서와의 유사도가 차이가 있는 지를 분석하기 위하여 ANOVA 분석을 시행하였으며, 결과는 <표 9>에 정리되어 있다. 연관도서 전체의 유사도를 도서 분야별로 비교하였으며, 세 가지 유사도 모두 도서 분야별로 유사도의 차이가 없는 것으로 분석되었다.

〈표 9〉 도서 분야별 유사도 비교

	F-value	P-value
Dice 유사도	1.656	0.197
Jaccard 유사도	1.984	0.144
Cosine 유사도	0.730	0.485

도서별로 연관도서가 연관도 1위부터 5위까지 존재하기 때문에 연관도서의 연관도 순위에 따라 유사도가 차이가 있는지를 비교하였다. ANOVA 분석을 수행하였으며, 연관순위간의 유사도 차이에 대해 Duncan 테스트를 통해 post hoc 분석을 실시하였다. 분석결과는 <표 10>에 정리되어 있다.

〈표 10〉 도서 연관도별 유사도 비교

도서 분야	유사도 방안	F value	P value	Duncan 테스트
비즈니스	Dice	2.271	0.064	1 > 4, 5
	Jaccard	2.031	0.093	1 > 4, 5
	Cosine	2.226	0.069	1 > 4, 5
문학	Dice	2.331	0.059	1 > 5
	Jaccard	2.782	0.029	1 > 3, 4, 5
	Cosine	3.429	0.010	1 > 3, 5
컴퓨터 네트워크	Dice	1.926	0.109	1 > 4
	Jaccard	1.814	0.129	1 > 4
	Cosine	1.773	0.137	1 > 4

ANOVA 분석을 통해 연관순위간의 차이가 있다고 볼 수 있는 경우는 문학의 Jaccard와 Cosine 유사도의 경우이다. 대체로 문학도서들이 연관도 순위에 따라 유사도의 차이가 존재하는 것으로 볼 수 있으며, 컴퓨터 네트워크의 경우에는 이러한 차이가 존재하지 않으며, 비즈니스 분야에는 신뢰수준 0.1에서 차이가 있는 것으로 보여 약간의 차이가 존재하는 것으로 볼 수 있다.

Duncan 테스트에서는 모든 경우에 연관도 1순위인 도서와 연관도가 5순위나 4순위인 도서 간에는 유사도의 차이가 존재하며, 연관도 1순위도서가 기준이 되는 도서와 더욱 유사한 것으로 볼 수 있다.

다음은 추천도서의 유사도와 연관도서의 유사도 간의 차이가 있는지를 분석하였다. 추천도서는 추천에 기반이 되는 도서와 추천된 도서가 한 쌍만 존재하기 때문에 연관도서와의 비교를 위해 연관순위가 1위인 도서와 해당 도서간의 유사도와 비교한 결과가 <표 11>이다. 추천도서와 분야별 연관도서의 유사도 평균값을 ANOVA 분석한 결과 P-value가 0.05보다 작아 평균 값이 다른 것으로 파악되었으며, Duncan 테스트 결과 에서 볼 수 있듯이 추천도서와 추천에 근거가 되는 도서간의 유사도가 분야별 베스트셀러 도서와 이의 연관순위 1위 도서와의 유사도 보다 작은 것으로 파악되었다. 추천도서와의 유사도와 분야별 베스트셀러 도서의 5개 연관도서와의 유사도 값을 비교하였으며, ANOVA 결과가 <표 12>에 정리되어 있다.

〈표 11〉 추천도서와 연관도 1위 도서간의 유사도 비교

유사도 방안	유사도	F-value	P-value	Duncan 테스트
Dice	0.185	4.708	0.004	추천도서 < 비즈니스, 문학, 컴퓨터 네트워크 연관도서
Jaccard	0.114	4.255	0.007	추천도서 < 비즈니스, 문학, 컴퓨터 네트워크 연관도서
Cosine	0.189	6.326	0.001	추천도서 < 비즈니스, 문학, 컴퓨터 네트워크 연관도서

연관도 1위 도서와 비교했을 때 보다는 P value가 높아졌지만, 여전히 그룹간 차이가 있는 것으로 분석되었다. Duncan 테스트 결과 유사도 평균이 비즈니스와 컴퓨터 네트워크에 비해 낮은 문학 도서 그룹은 추천도서와 유사하게 볼 수 있으며, 추천도서의 유사도는 비즈니스와 컴퓨터 네트워크 연관도서와의 유사도 보다 낮은 값을 보였다.

〈표 12〉 추천도서와 연관 도서간의 유사도 비교

유사도 방안	F value	P value	Duncan 테스트
Dice	2.959	0.035	추천도서 < 비즈니스
Jaccard	3.217	0.025	추천도서 < 비즈니스, 컴퓨터 네트워크 연관도서
Cosine	3.743	0.013	추천도서 < 비즈니스, 컴퓨터 네트워크 연관도서

추천도서가 대부분 비즈니스와 컴퓨터 분야에 해당하는 도서들이기 때문에 비즈니스와 컴퓨터 네트워크 분야의 연관순위 1~5위 연관도서의 유사도 평균값이 추천도서의 유사도 평균값보다 큰 것으로 볼 수 있다.

5. 토의 및 결론

본 연구에서는 구매 연관관계에 있는 상품들과 개인화 추천된 상품의 내용 유사도를 분석하였다. Amazon.com에서 비즈니스, 문학, 컴퓨터 네트워크 분야의 구매 연관도서 정보를 수집하여, 사용자들이 입력한 태그에 기반을 두어 도서간의 내용 유사도를 계산하였다.

대체로 비즈니스 분야와 컴퓨터 네트워크 분야의 내용 유사도가 높았으며, 문학 분야의 내용 유사도가 상대적으로 낮았다. 그러나 도서 분야별 유사도 평균을 통계적으로 비교하였을 때, 차이가 없는 것으로 분석되었다. 연관도서 전체의 내용 유사도 평균은 0.2에서 0.3수준이다. 한 도서가 평균적으로 8개에서 9개의 태그를 갖기 때문에 도서에 공통으로 존재하는 태그가 약 2개에서 3개 정도인 것으로 볼 수 있다. 구매 연관 순위에 따른 유사도는 연관순위가 높을수록 높은 유사도 값을 보였으나, 문학 분야에서는 순위에 따른 유사도 차이가 명확하게 존재하였으나 컴퓨터 네트워크와 비즈니스 분야는 명확한 차이는 없었다.

추천도서는 연관순위 1위 구매 연관도서와 비교하였을 때 유사도 값이 상대적으로 낮은 것으로 분석되었다. 구매연관도서 모두의 평균 유사도와 비교하였을 때는 문학 분야와 추천도서의 유사도가 비슷하지만, 다른 분야 보다는 낮은 유사도 값을 보였다.

태그를 활용하여 도서간의 유사도를 측정할 때 도서의 분야와 상관없이 비슷한 유사도 값을 보이기 때문에, 도서의 다양한 장르에서 태그를 활용하는 것이 가능한 것으로 볼 수 있다. 비즈니스와 문학 분야의 총 고유 태그 수가 638개, 501개로 409개의 컴퓨터 네트워크 분야 보다 많은 것에서 알 수 있듯이, 넓은 범위를 다루는 분야의 경우 일반적으로 태그가 다양해지는 경향이 있다. 하지만, 이런 경우에도 연관도서의 내용 유사도는 범위가 작은 경우와 비슷한 수준을 보인다. 물론, 통계적으로 차이는 없었지만 문학 분야보다는 비즈니스나 컴퓨터 네트워크의 유사도 값이 더 높았다. 따라서 태그에 기반을 둔 검색이나 유사도서 검색에는 분야에 따른 이러한 차이를 고려하는 것이 필요하다. 태그의 유사도가 높지 않은 분야에서는 메타데이터로 도서 정보나 상세페이지에 있는 다른 정보들을 추가적으로 활용하여야 할 필요가 있다.

개인화 추천도서는 구매 연관도서와 다르게 같은 분야의 도서에 국한되지 않으며, 구매 연관도서에서 많이 등장하는 동일 저자의 도서나 시리즈 도서가 추천되는 경우가 드물기 때문에 태그 유사도 값이 상대적으로 낮은 것으로 볼 수 있다. 그럼에도, 0.1에서 0.2정도의 유사도 값을 보이고 있다. 따라서 개인화 추천의 경우에도 추천의 기반이 되는 도서와 태그 일치도 면에서 어느 정도 유사도

를 보이는 책이 추천된 다는 것을 알 수 있다.

본 연구에서는 비즈니스, 문학, 컴퓨터 네트워크 분야의 도서정보만을 활용하였으나, 위 제품들과 성격이 다른 제품들의 정보를 활용하여 상관관계를 분석하여야 할 필요가 있다. 또한, 본 연구에서는 태그들의 의미를 고려하지 않고 서로 같은 태그가 존재하는 지만을 고려하여 유사도를 측정하였다. 정확히 일치하지 않는 태그라도 의미적 유사성을 가지고 있을 수 있기 때문에 이를 고려하여 유사도를 측정하여 결과를 분석하는 것이 필요하리라고 본다. 개인화 추천된 도서들이 저자의 성향을 고려하여 추천된 것이지만, 추천에 활용되는 알고리즘은 동일하기 때문에 추천도서의 내용 유사도는 개인 성향에 의한 차이가 그리 크지 않을 것이다.

따라서 태그를 활용하여 상품의 검색이나 연관상품을 제시하는 경우 해당 분야 상품간의 유사도가 얼마나 되는지, 태그들이 어떤 특성을 갖는지 고려하여 제공하는 것이 필요하다고 볼 수 있다.

참 고 문 헌

[1] 강상욱, 이기용, 김현규, 김명호, "태그를 이용한 웹 페이지간의 유사도 측정 방법", 「정보과학논문지 : 데이타베이스」, 제37권, 제2호(2010), pp.104-112.

[2] 김두남, 이강표, 김형주, "태그 동시 출현의 동적인 특징을 이용한 개선된 태그 클라우드의 태그 선택 방법", 「정보과학논문지 : 컴퓨팅의 실제 및 레터」, 제15권, 제6호(2009), pp.405-413.

[3] 김은희, 정영미, "사용자 태그와 중심성 지수를 이용한 블로그 검색 성능 향상에 관한 연구", 「정보관리학회지」, 제27권, 제1호(2010), pp.61-77.

[4] 김재경, 오희영, 권오병, "유비쿼터스 환경에서 연관규칙과 협업필터링을 이용한 상품그룹 추천", 「한국IT서비스학회지」, 제6권, 제2호(2007), pp.113-123.

[5] 김지혜, 박두순, "연관규칙과 협업적 필터링을 이용한 상품 추천 시스템 개발", 「한국컴퓨터교육학회논문지」, 제2권, 제2호(2006), pp.1-10.

[6] 김현우, 이강표, 김형주, "태깅 시스템의 태그 추천 알고리즘", 「정보과학논문지 : 컴퓨팅의 실제 및 레터」, 제16권, 제9호(2010), pp.927-935.

[7] 김형도, "잠재 요인 모델의 원리를 이용한 협업 태그 기반 추천 방법", 「한국전자거래학회지」, 제14권, 제4호(2009), pp.47-57.

[8] 박수진, 이시화, 황대훈, "소셜 북마킹 시스템에서의 북마크와 태그 정보를 활용한 웹 콘텐츠 랭킹 알고리즘", 「멀티미디어학회논문지」, 제13권, 제8호(2010), pp.1245-1255.

[9] 박수환, 김종우, 이홍주, 조남재, "전자상거래 개인화 추천을 위한 상품 카테고리 중립적 사용자 프로파일링", 「경영정보학연구」, 제16권, 제3호(2006), pp.143-160.

[10] 안현철, 한인구, 김경재, "연관규칙기법과 분류모형을 결합한 상품 추천 시스템 : G인터넷 쇼핑몰의 사례", 「Information Systems Review」, 제8권, 제1호(2006), pp.181-201.

[11] 엄태영, 김우주, 박상언, "태그 네트워크를 이용한 개인화 북마크 추천시스템", 「한국전자거래학회지」, 제15권, 제4호(2010), pp.181-195.

[12] 연철, 지애띠, 김홍남, 조근식, "효과적인 추천 시스템을 위한 협업적 태그 기반의 여과기법", 「지능정보연구」, 제14권, 제2호(2008), pp.157-177.

[13] 이홍주, "소셜 태깅 사이트에서의 제시 태그 개수 예측에 대한 연구", 「산업경영연구」, 제16호(2008), pp.85-98.

[14] 이홍주, "클릭스트림 데이터를 활용한 전자상거래에서 상품추천이 고객 행동에 미치는 영향 분석", 「한국경영과학회지」, 제33권, 제3호(2008), pp.61-78.

[15] 황인수, "연관규칙을 이용한 상품선택과 기대수익 예측", 「경영정보학연구」, 제14권, 제4호

(2004), pp.87-97.

[16] 황현숙, 어윤양, "연관 마이닝과 고객 선호도 기반의 인터넷 상품 검색 시스템 설계 및 구현", 「경영정보학연구」, 제12권, 제1호(2002), pp.1-16.

[17] Frank, E. and I. H. Witten, Data Mining : Practical machine learning tools and techniques, Morgan Kaufman, 2nd ed., 2005.

[18] Golder, S. A. and B. A. Huberman, "Usage patterns of collaborative tagging systems", *Journal of Information Science*, Vol.32, No.2 (2006), pp.198-208.

[19] Halpin, H., V. Robu, and H. Shepherd, "The Complex Dynamics of Collaborative Tagging", WWW 2007 : Proceeding of the 16th international conference on World Wide Web, (2007), pp.211-220.

[20] Jiao, Y. and G. Cao, "A Collaborative Tagging System for Personalized Recommendation in B2C Electronic Commerce", Proceeding of the International Conference on Wireless Communications, Networking and Mobile Computing, (2007), pp. 3609-3612.

[21] Linden, G., B. Smith, and J. York, "Amazon.com Recommendations : Item-to-item Collaborative Filtering", *IEEE Internet Computing*, Vol.7, No.3(2003), pp.76-80.

[22] McAfee, A., Enterprise 2.0 : New Collaborative Tools for Your Organization's Toughest Challenges, Harvard Business School Press, Cambridge : MA, 2009.

[23] Penev, A. and R. Wong, "Finding similar pages in a social tagging repository", WWW 2008 : Proceeding of the 17th international conference on World Wide Web, (2008), pp. 1091-1092.

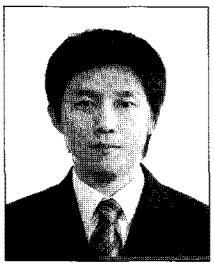
[24] Rhie, B. W., J. W. Kim, and H. J. Lee, "Methods of User-Created Content Recommendation with Content Metadata", *International Journal of Management Science*, Vol. 16, No.2(2010), pp.29-38.

[25] Sarwar, B., G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms", Proceedings of the 10th International Conference on World Wide Web, (2001), pp.285-295.

[26] Senecal, S., P. Kalczynski, and J. Nantel, "Consumers' decision-making process and their online shopping behavior : a clickstream analysis", *Journal of Business Research*, Vol.58, No.11(2005), pp.1599-1608.

[27] Sigurbjörnsson, B. and R. Zwol, "Flickr tag recommendation based on collective knowledge", WWW 2008 : Proceeding of the 17th international conference on World Wide Web, 2008, pp.327-336.

◈ 저 자 소 개 ◈



이 홍 주 (hongjoo@catholic.ac.kr)

KAIST 산업경영학과를 졸업하고 KAIST 경영대학원에서 경영정보시스템으로 석사와 박사학위를 받았다. 현재 가톨릭대학교 경영학부 교수로 재직중이며 주요 연구 관심분야는 웹 개인화, 온라인 협업, 인공지능 기법 응용이다.