DOI : 10.5392/JKCA.10.10.009

디지털 콘텐츠를 위한 소속도를 이용한 사례기반 필터링

Case-based Filtering by Using Degree of Membership for Digital Contents

김형일
나사렛대학교 멀티미디어학과

Hyungil Kim(hkim@kornu.ac.kr)

요약

디지털 콘텐츠의 양이 방대해지면서 사용자가 원하는 디지털 콘텐츠를 검색하는 데 많은 시간이 필요하다. 그러므로 방대한 디지털 콘텐츠로부터 사용자가 원하는 콘텐츠를 제공하기 위해서는 디지털 콘텐츠를 분석하여 사용자에게 적합한 콘텐츠를 추출하는 기술이 필요하다. 그리고 빠른 시간 내에 사용자에게 적합한 디지털 콘텐츠를 찾기 위해서는 디지털 콘텐츠에 대한 필터링 기술이 필요하다. 본 논문에서는 개인에게 적합한 디지털 콘텐츠를 필터링하는 기법을 제안한다. 본 논문에서 제안한 기법은 디지털 콘텐츠에 대한 사례기반 정보를 분석하여 개인 사용자에게 적합한 디지털 콘텐츠를 제공한다. 사용자의 선호도 분석에는 디지털 콘텐츠 사용에 대한 사례를 이용한다. 다양한 시뮬레이션을 통해 제안한 기법의 효과를 확인하였다.

■ 중심어 : | 디지털 콘텐츠 | 정보여과 | 협업여과 | 개인화 |

Abstract

As digital contents become vast in quantity, it takes long time for users to search the digital contents they want, which is a problem that has arisen. Therefore, it is required to have the technology that analyzes vast digital contents and extracts the appropriate contents for users in order to provide them with contents they want. For a fast searching of digital contents suitable for users, it is necessary to have the technology of filtering for digital contents. In this paper, we propose a method of filtering digital contents suitable for individual users. The method suggested in this paper is to analyze case-based information in digital contents and provide the digital contents suitable for individual users. The case for using digital contents is used for analysis of users' preference. Various simulations were conducted to confirm the effectiveness of the proposed method.

■ keyword : | Digital Contents | Information Filtering | Collaborative Filtering | Personalization |

I. 서론

디지털 콘텐츠는 소프트웨어 형태로 존재하기 때문에 무한 복제가 가능하고, 정보의 결합과 변형이 용이하여 적은 비용으로 부가가치를 극대화시킬 수 있다는 장점이 있다. 디지털 콘텐츠는 정보 갱신의 속도가 빨라서 결함이 존재하더라도 빠른 시간 내에 데이터를 갱신하여 결함을 보정할 수 있다. 정보통신 기기의 발달과 대용량 저장 매체의 보급으로 디지털 콘텐츠의 보관과 활용이 용이해졌으며, 네트워크를 통해 국내뿐만 아니라

* 본 논문은 2010년도 나사렛대학교 학술연구비 지원에 의해서 연구되었음.
접수번호 : #101004-003
접수일자 : 2010년 10월 04일
심사완료일 : 2010년 10월 25일
교신저자 : 김형일, e-mail : hkim@kornu.ac.kr

글로벌 시장으로의 접근성이 용이하다. 그리고 디지털 콘텐츠는 소프트웨어 형태로 존재하며, 네트워크를 통하여 이동하는 경우가 주류를 이루기 때문에 유통에 소요되는 시간과 비용이 매우 적다는 장점이 있다[1].

디지털 콘텐츠 산업분야에서 국내의 경우 2010년 2분기 월평균 오락 및 문화 부분의 지출이 전년도 대비 11.6% 증가하였으며, 2005년을 기준으로 2008년과 2009년의 연간 콘텐츠 산업생산지수는 115.5(2.6%)와 120.0(3.9%)로 증가하였다. 전자출판의 경우 정부가 2010년 "전자출판산업 육성방안"을 발표하였으며, 향후 5년간 600억 원을 투자하기로 결정한 사항을 고려한다면 디지털 콘텐츠의 빠른 발전이 예상된다. 또한 공공도서관에서는 전자책 구입을 확대할 예정에 있으며, 국내 전자책 시장규모는 2009년도에 1,200억 원 정도였고, 매년 20~30% 이상 성장할 것으로 예측하고 있다[2].

스마트폰은 통화기능뿐만 아니라 일반적인 PC에서 활용하는 기능을 일부 흡수한 다기능 개인 단말기이다. 스마트폰 제작사들은 스마트폰에서 활용되는 어플리케이션을 개인 개발자들도 개발할 수 있도록 허용함으로써 디지털 콘텐츠 개발의 활성화에 기여하고 있다.

전술한 바와 같이 디지털 콘텐츠의 활성화는 개인 사용자들에게 폭넓은 디지털 콘텐츠 활용 환경을 제공하였다. 그러나 디지털 콘텐츠의 기하급수적인 증가는 사용자가 원하는 디지털 콘텐츠를 검색하는 데 많은 시간이 필요하다는 문제를 발생시킨다. 그러므로 방대한 디지털 콘텐츠로부터 사용자가 원하는 콘텐츠를 효과적으로 빠른 시간 내에 추출하기 위해서는 디지털 콘텐츠를 분류하고 분석하는 기술이 시급히 필요하다. 다기능 개인 단말기의 확산은 디지털 콘텐츠 사용에 있어서 공통된 주제로의 접근보다는 개인화된 주제로의 접근을 요구한다.

본 논문에서는 방대한 디지털 콘텐츠에서 정보 사용자의 사용 사례를 분석하여 사용자에게 적합한 디지털 콘텐츠를 추출하는 기법을 제안한다. 본 논문에서 제안한 기법은 사용정보가 잘 표현되지 않은 디지털 콘텐츠에서도 효과적으로 콘텐츠를 필터링할 수 있는 기법이다.

본 논문의 2장에서는 관련 연구에 대해 설명하고, 3장에서는 본 논문에서 제안한 소속도를 이용한 사례기반 필터링에 대해 설명한다. 4장에서는 제안한 기법의 시뮬레이션 결과를 분석하고, 5장에서는 결론과 향후 연구에 대해 기술한다.

II. 관련 연구

디지털 콘텐츠의 양이 방대해짐에 따라 디지털 콘텐츠를 효과적으로 분류, 저장, 검색, 활용하기 위해서는 디지털 콘텐츠의 활용성과 상호 호환성을 증대시키기 위한 연구가 필요하다.

이러한 연구 중 한 가지가 국제 표준기구인 ISO/IEC에서 추진하는 온톨로지(ontology) 기술표준인 토픽맵(topic map)이다. 온톨로지는 서로 다른 형태로 존재하는 객체들에 의미를 부여하여 공유 객체로 활용할 수 있는 방법을 제공한다. 토픽맵은 지식을 표현하고, 관련 정보자원과 연계하기 위한 추상적 구조이다[3][4]. 토픽맵은 지식층과 정보층으로 구성되며, 토픽(topic), 연관(association), 발생(occurrence)의 세 가지 요소를 사용하여 정보를 표현한다. 지식층은 토픽들에 대한 지식구조를 표현한 영역이며, 정보층은 지식에 대한 자원영역으로 자원과 자원에 대한 정보를 담고 있는 메타데이터로 구성된다. 토픽맵을 구성하는 요소 중 토픽은 지식구조로 표현할 개념이다. 연관은 개념 간의 관계성을 표현한 것이며, 발생은 개념과 정보자원과의 관계를 표현한 것이다. Park와 Cheyer[5]는 데스크탑에서 토픽맵을 활용하는 IRIS 플랫폼을 개발하였다. IRIS 사용자들은 토픽맵을 활용함으로 개인화된 지식층을 구축하고 활용할 수 있다. Seedorf 등[6]은 모바일 단말기에서 토픽맵 서버에 접속하여 토픽맵을 검색하고 활용하는 토픽맵 질의 도구를 자바를 사용하여 개발하였으며, PDA에서 활용할 수 있게 설계되었다. 그리고 XML 데이터를 사용하여 토픽맵 정보를 표현하고 활용하였다. 그러나 이와 같은 기술들은 상용화되면 디지털 콘텐츠에서 효과적으로 사용될 수 있으나, 아직 상용화하기에는 추가적인 연구와 시간이 필요하다.

생산과 유통에 대한 저비용과 활용의 편리함 때문에 전자책은 최근 많은 활성화를 이룬 분야로 전자책 연구

는 전자책을 읽는 환경이나 전자책 인터페이스를 중심으로 진행되고 있다[7][8]. Card 등[9][10]은 3D 전자책인 3Book을 설계 및 구현하였다. 3Book은 3D 가상환경에서 주석 달기, 밑줄 긋기 등과 같은 다양한 기능을 제공하며, 현실적인 책장 넘김 효과를 표현하였다. Chu 등[11]은 자연스러운 전자책 읽기를 제공하기 위해 페이지 넘김 방법에 대해 연구하였으며, 결과물은 영국국립도서관의 디지털북에 적용되었다. Agoshkov 등[12]은 전자출판의 모든 과정을 통합 처리하고 관리하는 전자출판 관리시스템을 제안하고, Elysium라는 전자출판 관리시스템을 개발하였다. Elysium은 준비, 로딩, 뷰어 및 외부 상호작용 단계로 구성된다. 이러한 통합된 전자출판 관리시스템은 전자출판 처리 과정의 편리성과 개발의 효율성을 제공한다. 이러한 전자책 환경과 인터페이스 기술들은 많은 연구가 진행되고 있지만, 기하급수적으로 증가되는 디지털 콘텐츠에서 사용자에게 적합한 콘텐츠를 추출하는 연구는 미진한 상태이다. 방대한 수량의 디지털 콘텐츠를 고려한다면 사용자에게 적합한 콘텐츠 추출에 대한 심도 있는 연구가 필요하다.

사용자에게 적합한 디지털 콘텐츠를 추출하기 위해서는 디지털 콘텐츠 분석이 필요하며, 디지털 콘텐츠 내용을 분석하는 것이 콘텐츠 분석에 매우 효과적이다. 그러나 디지털 콘텐츠 저작자들은 불법 사용에 대응하기 위해 다양한 방법으로 내용정보 확인을 제한한다. 또한 콘텐츠 저작자의 권리를 보호하기 위해 DRM(Digital Rights Management) 기술을 콘텐츠에 적용하기 때문에 임의 사용자가 DRM을 해제하여 콘텐츠의 내용을 사용할 수 없는 현실을 고려할 때 내용기반 기술로 디지털 콘텐츠의 정보를 분석한다는 것은 매우 어려운 일이다. 이러한 상황에서 방대한 디지털 콘텐츠를 대상으로 사용자에게 적합한 콘텐츠를 추출한다는 것은 매우 어려운 작업에 속한다. 사용자의 성향 예측에 대한 사용되는 대표적인 기법은 협력적 여과이며, 협력적 여과는 사용자 성향을 예측할 때 유사 사용자들의 성향을 이용한다. GroupLens[13] 연구에서는 사용자의 성향 예측에 상관관계를 사용하였고, 다양한 변형 기법들이 정확도 향상을 위해 제안되었다. Breese [14], Herlocker 등[15]은 다양한 유사도 계산과 유사도 가중치에 대한 실험을 수행한 바 있으며, Billsus와 Pazzani[16], Sarwar 등[17]은 충분한 정보가 없을 때 사용자의 선호도 예측을 위해 속성 추출을 적용하는 방법과 SVD(Singular Value Decomposition)를 이용하여 사용자-아이템 선호도 행렬의 차원을 줄이는 방법 등을 제안하였다. 사용자 사이의 유사도를 비교하는 방법 이외에 Deshpande[18], Sarwar[19], Linden[20] 등에 의해 아이템 사이의 유사도를 이용하여 사용자의 증가에 따른 계산 복잡도 문제를 해결하면서 추천의 질을 높일 수 있는 아이템기반(item-based) 협력적 여과 알고리즘이 제안되었다. 협력적 여과는 사용자 성향을 예측할 때 유사 사용자의 성향을 이용하기 때문에 사용자 성향 파악에 효과적으로 적용될 수 있는 기법이지만, 디지털 콘텐츠와 같이 대량의 집합에서 사용자의 사용정보가 매우 희박하게 나타날 경우에는 성능이 급격히 떨어지는 문제가 있다.

본 논문에서는 이와 같은 방대한 디지털 콘텐츠에서 사용자에게 적합한 콘텐츠를 효과적으로 추출하기 위해 소속도를 이용한 사례기반 필터링을 제안한다. 본 논문에서 제안한 기법은 디지털 콘텐츠에 접근한 사용자들의 묵시적 행위 분석을 통해 사용자의 콘텐츠 사용 사례에 기반하여 콘텐츠 일반 속성을 확장하고, 이러한 과정을 통해 콘텐츠 분석에 활용되는 정보량을 증가시켜 사용자 성향 분석에 이용한 후, 사용자에게 적합한 콘텐츠를 추출한다. 본 논문에서 제안한 기법은 내용기반 분석이 불가능한 디지털 콘텐츠 데이터에서도 사용자에게 적합한 콘텐츠를 효과적으로 추출할 수 있으며, 사용자의 성향 분석에 사용되는 정보가 희박한 경우에도 효과적으로 사용자 성향을 분석하여 디지털 콘텐츠를 추출할 수 있는 장점이 있다.

III. 소속도를 이용한 필터링

디지털 콘텐츠를 분석하기 위해서 내용기반기법을 적용하게 되면 매우 효과적일 수 있으나, 일반적으로 디지털 콘텐츠는 저작권에 의해 내용을 공개하지 않기 때문에 내용기반기법으로 디지털 콘텐츠를 분석할 수

없다. 이러한 상황에서 디지털 콘텐츠를 분석하기 위해 정보 사용자에게 능동적 행위(사용기, 평가 등)를 요구하여 명시적 정보를 디지털 콘텐츠 분석에 활용하면 효과적일 수 있다. 이와 같은 명시적 정보는 디지털 콘텐츠 분석에 중요한 정보이나, 일반적으로 정보 사용자들은 자신의 능동적 행위를 나타내지 않으려는 성향이 있어 사용자의 명시적 정보를 취득하기란 매우 어렵다.

이와 같은 상황에서는 사용자의 묵시적 행위를 정보화시켜야 하며, 대표적인 방법이 정보 접근에 대한 정보화이다. 예를 들면 디지털 콘텐츠 사용을 위해 사용자들은 스트리밍 서비스나 다운로드 서비스 등을 받게 될 것이며, 이와 같은 묵시적인 행위를 정보화시켜 사용 성향 분석에 활용할 수 있다. 그러나 이와 같은 묵시적 정보는 불리언 형식으로 저장되는 한계가 있으며, 방대한 디지털 콘텐츠 수량에 비해 사용자가 접근한 디지털 콘텐츠의 수량은 매우 적다는 문제도 있기 때문에 성향 분석을 어렵게 만든다.

취득한 묵시적 정보(사용정보)를 이용하여 사용자에게 적합한 디지털 콘텐츠를 제공할 때는 정보 부족 문제로 우수한 결과를 생성하기가 매우 어렵다. 정보 부족과 같은 문제를 해결하기 위해 협력적 여과를 적용하기도 하나, 디지털 콘텐츠와 같은 방대한 크기의 영역에서는 정보 부족이 매우 심각하여 성향 분석이 어렵다는 것이 문제이다. 확률에 기반한 성향 분석도 많이 활용되기는 하나, 정보 부족의 문제로 확률에 기반한 성향 분석도 방대한 크기의 영역에서는 사용이 제한적이다.

기본적으로 묵시적 사용정보만을 활용한다면 사용성향을 측정하기란 매우 제한적이며, 성향 예측값 또한 불순도가 매우 높을 수 있다. 이와 같은 문제를 완화하기 위해 본 논문에서는 소속도를 이용한 사례기반 필터링 방법을 제안한다.

일반적으로 디지털 콘텐츠는 분야, 심의등급 등과 같은 일반 속성을 포함하며, 일반 속성은 다양한 체계로 디지털 콘텐츠를 나누게 된다. 이와 같은 디지털 콘텐츠의 일반 속성을 이용하면 디지털 콘텐츠의 필터링에 효과적일 수 있다. 본 논문에서 제안한 방법은 디지털 콘텐츠의 일반 속성의 확장 적용을 위해 정보 사용자의 접근 사례에 기반한 소속도를 적용하였다. 접근 사례에 기반한 소속도를 이용하여 일반 속성을 확장한 후, 디지털 콘텐의 속성으로 활용하면 정보량을 풍부하게 하여 사용자에게 적합한 디지털 콘텐츠를 제공할 수 있다는 장점이 있다.

임의 사용자가 접근하지 않았던 디지털 콘텐츠에 대한 성향 예측에 성향이 유사한 사용자들을 이용하면 매우 효과적이다. 사용자 유사도를 측정하는 방법은 여러 가지가 있으며, 대표적인 방법은 벡터 유사도와 피어슨 상관관계이다. 피어슨 상관관계를 이용한 유사도 측정을 식 1에 나타내었다.

$$S_{u,k} = \frac{\sum_{i=1}^{n}(r_{u,i} - \overline{r_u}) \bullet (r_{k,i} - \overline{r_k})}{\sigma_u \bullet \sigma_k} \quad (1)$$

$S_{u,k}$는 사용자 u와 사용자 k의 유사도이다. $r_{u,i}$는 사용자 u의 콘텐츠 i에 대한 선호도이고, $r_{k,i}$는 사용자 k의 콘텐츠 i에 대한 선호도이다. $\overline{r_u}$는 사용자 u의 평균 선호도를 나타내고, $\overline{r_k}$는 사용자 k의 평균 선호도를 나타낸다. σ_u는 사용자 u의 선호도 표준편차이고, σ_k는 사용자 k의 선호도 표준편차이다.

사용자가 접근한 경험이 없는 디지털 콘텐츠에 대한 선호도를 예측할 경우에는 식 2에서와 같이 디지털 콘텐츠에 대해 유사한 사용자들의 선호도를 활용하여 개인 사용자의 콘텐츠 선호도를 예측한다. $P_{u,i}$는 사용자 u의 콘텐츠 i에 대한 선호도 예측값이다. $r_{k,i}$는 사용자 k의 콘텐츠 i에 대한 선호도이다. $S_{u,k}$는 사용자 u와 사용자 k의 유사도이다. $\overline{r_u}$는 사용자 u의 평균 선호도를 나타내고, $\overline{r_k}$는 사용자 k의 평균 선호도를 나타낸다.

$$P_{u,i} = \overline{r_u} + \frac{\sum_{k=1}^{n}(r_{k,i} - \overline{r_k}) \bullet S_{u,k}}{\sum_{k=1}^{n} S_{u,k}} \quad (2)$$

디지털 콘텐츠의 일반 속성을 확장하기 위한 소속도 생성은 일반 속성에 나타난 분류 영역을 이용한다. 이러한 분류 영역을 활용하여 일반 속성을 재표현한다. 예를 들어 임의의 콘텐츠를 x라 하고 분류 영역을 c_1, c_2,...,c_i, ...,c_n이라 가정하면 임의의 콘텐츠 x는 n개의 임의의 분류 영역 중에 속하게 된다. 그러나 콘텐츠는 이러한 분류 영역에 다시 계층화 되는 경우가 발생하기도 한다. 대표 분류를 결정하기 위해 각 영역에 대한 소속도를 측정한 후에 대표 소속도를 생성하여 콘텐츠의 대표 영역을 추출한다. 소속도를 측정할 경우에는 정보 사용자의 사용 사례에 기반하여 계산하기 때문에 콘텐츠의 영역 대표성을 나타낼 수 있어 인기도를 측정할 수 있다는 장점이 있다.

사례기반 소속도 측정 방법을 식 3에 표현하였다. $D_{f_i}(x)$는 임의의 콘텐츠 x가 대표 영역 f_i에 속하는 대표 영역 소속도를 나타낸다. $P(c_i|x)$는 콘텐츠 x가 영역 c_i에 속하는 단일 영역 소속도를 나타낸다. 이와 같은 방법을 모든 콘텐츠에 적용하여 단일 영역 소속도와 대표 영역 소속도를 측정하여 콘텐츠의 속성 확장에 적용한다. 영역 소속도는 임의의 디지털 콘텐츠 x가 소유한 속성 벡터를 활용하여 계산할 수 있다. 속성 벡터는 임의의 콘텐츠가 나타내는 일반 정보로써 콘텐츠에 나타난 분야나 심의등급 등을 활용하며, 정상적으로 배포된 대다수의 디지털 콘텐츠들은 이와 같은 디지털 콘텐츠 일반 정보를 소유한다. 콘텐츠의 일반 정보를 속성 벡터로 활용하여 단일 영역 소속도를 계산할 수 있다. 식 3을 콘텐츠의 속성 원소를 이용하여 재표현하면 식 4와 같다. 식 4에 나타난 $a_1, a_2, \ldots, a_n$는 디지털 콘텐츠의 속성 원소가 된다.

$$D_{f_i}(x) = Max(P(c_1|x), P(c_2|x), \ldots, P(c_i|x), \ldots, P(c_n|x)), \quad i,n \in N \tag{3}$$

$$= Max(P(c_1|a_1,a_2,\ldots,a_n), P(c_2|a_1,a_2,\ldots,a_n), \ldots, P(c_i|a_1,a_2,\ldots,a_n), \ldots, P(c_n|a_1,a_2,\ldots,a_n)) \tag{4}$$

본 논문에서 제안한 소속도를 이용한 사례기반 필터링은 대량의 디지털 콘텐츠에 나타난 희소한 사용정보를 풍부하게 만드는 장점이 있다. 이와 같은 방법을 디지털 콘텐츠에 적용하면 정보 사용자의 콘텐츠 접근 성향을 파악할 수 있어 사용자에게 적합한 디지털 콘텐츠를 추출할 수 있다는 장점이 있다.

IV. 시뮬레이션

시뮬레이션은 2.0GHz Quad-Core CPU, 2Gb RAM, Windows XP 환경에서 수행하였다. 성능 측정을 위해 시뮬레이션 프로그램을 이용하였고, 시뮬레이션 프로그램은 Visual C++ 6.0을 이용하여 개발하였으며, [그림 1]에 개발한 시뮬레이션 프로그램을 표현하였다.

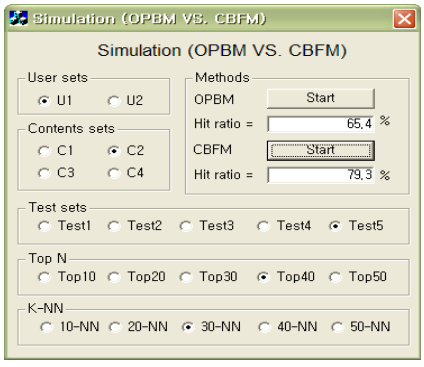

그림 1. 시뮬레이션 프로그램

시뮬레이션을 위해 시뮬레이션 데이터를 생성하였으며, 시뮬레이션 데이터 C1과 C2는 디지털 콘텐츠가 1,000개와 2,000개로 구성되며, 시뮬레이션 데이터 C3, C4는 디지털 콘텐츠가 3,000개 5,000개로 구성된다. 시뮬레이션 데이터에 대한 사용자 집단의 크기는 U1과 U2로 나뉘며, 각 집단에 포함된 사용자 수는 1,000명과 2,000명이다. 디지털 콘텐츠 데이터와 사용자 데이터를 서로 결합하여 최종 시뮬레이션 데이터를 생성할 때는 다음과 같은 방법을 따른다. C1과 U1의 결합으로 D11 데이터를 생성하며, 생성된 최종 데이터의 크기는 1,000X1,000이다. C1과 U2의 결합으로 D12 데이터를 생성하며, 생성된 최종 데이터의 크기는 1,000X2,000이

다. 이와 같은 방법으로 D21, D22, D31, D32, D41, D42 데이터를 생성한다. 일반적으로 디지털 콘텐츠는 분야, 심의등급 등과 같은 공개된 정보를 제공하며, 이와 같은 현실적 상황을 고려하여 모든 시뮬레이션 데이터를 그룹화하였다. 각 디지털 콘텐츠 데이터의 그룹은 20개로 제한하였으며, 정보 사용자의 묵시적 정보는 일반적으로 접근 정보에서 추출하기 때문에 불리언 형식을 이용하여 표현하였다. 그러므로 사용정보 데이터에 나타난 '1'은 콘텐츠 접근(예 : 다운로드나 스트리밍 서비스 등)을 의미하고, '0'은 해당 콘텐츠에 접근하지 않은 것을 의미한다. 디지털콘텐츠에 대한 사용자의 사용정보를 생성할 때는 사용정보 희소성을 표현하기 위해 디지털콘텐츠 그룹별로 5%~10% 범위에서 랜덤하게 생성하였다.

콘텐츠 추출기법에 따른 콘텐츠 추출에는 TopN을 적용하였으며, N은 10, 20, 30, 40, 50로 한정하였다. 사용자 유사도 측정에는 K-Nearest Neighbor를 적용하였으며, 유사 사용자 추출을 위한 K는 10, 20, 30, 40, 50으로 제한하였다. K값에 따른 추출 성능은 산술평균을 적용하여 대표 성능을 생성하였고, 시뮬레이션에 적용된 디지털 콘텐츠 추출기법은 일반적으로 많이 활용하는 콘텐츠 발생에 기원한 발생확률기반 추출기법(Occurrence Probability-Based Method, OPBM)과 본 논문에서 제안한 사례기반 필터링기법(Case-Based Filtering Method, CBFM)이며, 시뮬레이션 수행에는 5-fold cross validation을 적용하였다. 성능 측정을 수행 때는 테스트 사용자가 접근한 콘텐츠 사용정보를 삭제하고, 추출기법에 따른 결과를 생성하여 삭제된 사용정보와 동일한가를 측정한다. 삭제된 사용정보와 추출기법의 결과가 동일하면 적중한 것이고, 동일하지 않으면 실패한 것으로 결정한다. 추출기법에 대한 성능 측정에는 적중률을 사용하였으며, 적중률을 식 5에 나타내었다.

$$Hit\ ratio = \frac{total\ number\ of\ hits}{total\ number\ of\ tests} \quad (5)$$

[그림 2]는 D11 데이터를 이용한 시뮬레이션 결과이다. 콘텐츠 추출이 10개일 경우 OPBM은 45%의 적중률을 나타내었고, CBFM은 57%의 적중률을 나타내어 OPBM보다 12%의 성능 향상을 보였으며, 콘텐츠 추출이 20개일 경우 OPBM은 48%의 적중률을 나타내었고, CBFM은 61%의 적중률을 나타내었다. 콘텐츠 추출이 30개일 경우 OPBM은 52%의 적중률을 나타내었고, CBFM은 75%의 적중률을 나타내어 OPBM보다 23%의 성능 향상을 보였으며, 콘텐츠 추출이 40개와 50개일 경우 OPBM은 각각 65%와 72%의 적중률을 나타내었고, CBFM은 각각 79%와 85%의 적중률을 나타내었다. OPBM의 평균 적중률은 56%를 나타내었으며, CBFM의 평균 적중률은 71%를 나타내어 OPBM보다 평균 15%의 성능 향상을 보였다. D11은 시뮬레이션 데이터에서 두 번째로 사용정보가 많이 나타난 데이터이기 때문에 OPBM과 CBFM 모두 우수한 성능을 나타내었으며, 사용정보가 많이 나타나게 되면 사용자의 성향 예측 활용되는 정보가 많기 때문에 우수한 성능을 나타내는 것이다.

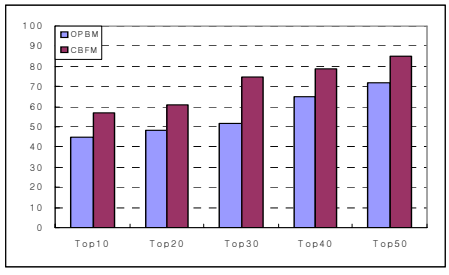

그림 2. D11 데이터의 결과

[그림 3]은 D12 데이터를 이용한 시뮬레이션 결과이다. 콘텐츠 추출이 10개일 경우 OPBM과 CBFM은 각각 42%와 64%의 적중률을 나타내었고, 콘텐츠 추출이 20개일 경우 OPBM과 CBFM은 각각 51%와 72%의 적중률을 나타내었으며, 콘텐츠 추출이 30개일 경우 OPBM과 CBFM은 각각 58%와 81%의 적중률을 나타내었다. 콘텐츠 추출이 40개와 50개일 경우 OPBM은 71%의 평균 적중률을 나타내었고, CBFM은 87%의 평균 적중률을 나타내었다. 적은 수량의 콘텐츠를 추출하여도 CBFM은 OPBM보다 우수한 성능을 나타내었으며, 적은 수량의 콘텐츠 추출에도 CBFM이 우수한 성

능을 나타내는 이유는 CBFM은 성향 예측에 유사 사용자들의 사용정보를 활용하여 희소한 사용정보를 보정하기 때문이다. 임의 사용자에게 유사 사용자의 사용정보를 활용하면 임의 사용자에게 나타난 희소한 사용정보를 완화하는 효과와 함께 임의 사용자에게 새로운 정보를 제공하는 효과가 있다. 이와 같은 유사 사용자의 정보 보정 효과를 본 시뮬레이션에 확인하였다.

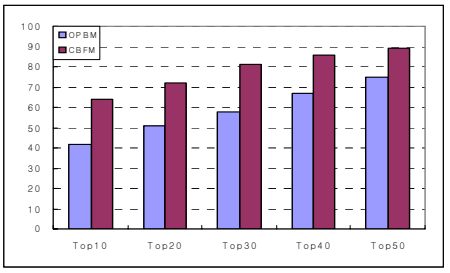

그림 3. D12 데이터의 결과

[그림 4]는 D21 데이터를 활용한 시뮬레이션 결과로 콘텐츠 추출이 10개일 경우 OPBM과 CBFM은 각각 35%와 44%의 적중률을 나타내었고, 콘텐츠 추출이 20개와 30개일 경우 OPBM은 각각 43%와 47%의 적중률을 나타내었고, CBFM은 각각 58%와 61%의 적중률을 나타내었다. 콘텐츠 추출이 40개와 50개일 경우 OPBM은 각각 62%와 68%의 적중률을 나타내었고, CBFM은 각각 72%와 75%의 적중률을 나타내었다.

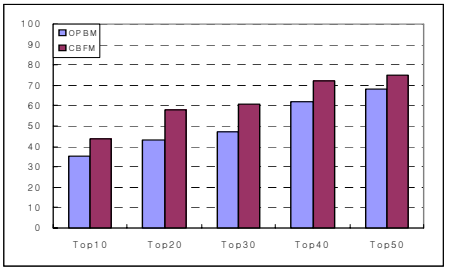

그림 4. D21 데이터의 결과

D21 데이터는 D11 데이터보다 사용정보가 적게 나타난 데이터이기 때문에 OPBM과 CBFM 모두 D21에서의 성능이 D11에서의 성능보다 낮게 나타났다. 이와 같은 결과가 발생한 이유는 대량의 디지털 콘텐츠보다 사용정보가 적게 나타나면 사용자들의 성향을 분석하는데 한계가 있기 때문이다. 이와 같은 정보 부족 문제를 해결하기 위해서는 정보량을 늘려야 하며, 본 논문에서 제안한 기법은 정보 희박성 문제를 완화하여 콘텐츠 추출에 적용하기 때문에 희소한 사용정보를 활용하더라도 높은 성능을 나타내는 것이다.

[그림 5]는 D22 데이터를 이용한 시뮬레이션 결과이다. 콘텐츠 추출이 10개와 20개일 경우 OPBM은 각각 41%와 47%의 적중률을 나타내었고, CBFM은 각각 58%와 65%의 적중률을 나타내었으며, 콘텐츠 추출이 30개일 경우 OPBM과 CBFM은 각각 52%와 69%의 적중률을 나타내었다. 콘텐츠 추출이 40개와 50개일 경우 OPBM은 각각 59%와 65%의 적중률을 나타내었고, CBFM은 각각 72%와 75%의 적중률을 나타내어 OPBM보다 각각 13%와 10%의 성능 향상을 나타내었다. 추출 콘텐츠의 수량이 증가함에 따라 CBFM의 OPBM에 대한 성능 향상이 줄어드는 이유는 사용자가 선호하는 콘텐츠가 OPBM의 경우 우선 순위가 낮게 책정되기 때문이다. 높게 나타나야 하는 디지털콘텐츠 우선 순위가 낮게 나타나는 이유는 개인의 성향 파악이 올바르게 수행되지 않기 때문이다. 발생확률에 기반한 콘텐츠 추출은 공통된 특성에 관련한 대량의 콘텐츠 추출에는 효과가 있지만, 개인에게 적합한 소량의 콘텐츠 추출에는 비효율적이라는 것을 여러 시뮬레이션 결과에서 확인할 수 있었다.

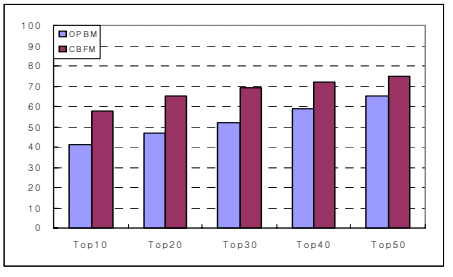

그림 5. D22 데이터의 결과

[그림 6]은 D31 데이터를 활용한 시뮬레이션 결과이고 [그림 7]은 D32 데이터를 활용한 시뮬레이션 결과이

다. D31에서의 OPBM의 평균 적중률은 36%로 나타났으며, CBFM의 평균 적중률은 48%로 나타났다. D31 데이터는 D21과 D11 데이터보다 사용정보가 적게 나타난 데이터인 경우이기 때문에 OPBM과 CBFM 모두 D31에서의 성능이 D21과 D11에서의 성능보다 낮게 나타났다. 디지털 콘텐츠에 대한 접근 사용자들의 수가 상황확률기반 추출기법이든 사례기반 필터링기법이든 중요한 변수로 작용한다는 것을 여러 시뮬레이션 결과로 확인할 수 있었다. D32 데이터는 D31 데이터보다 사용자의 수가 많은 데이터이다. D32에서 OPBM은 38%의 평균 적중률을 나타내었고, CBFM은 52%의 평균 적중률을 나타내었다. D31과 D32의 결과를 비교할 때 OPBM의 성능 향상보다 CBFM의 성능 향상이 더 높게 나타났다. 이와 같은 시뮬레이션 결과를 보더라도 CBFM은 OPBM에 비해 정보 활용도가 높다는 것을 알 수 있다.

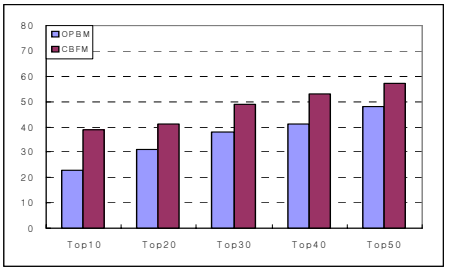

그림 6. D31 데이터의 결과

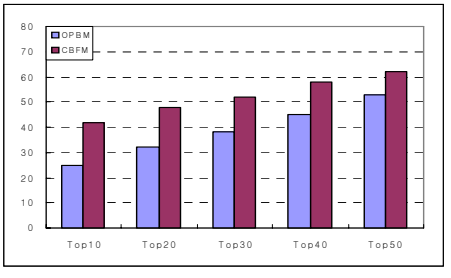

그림 7. D32 데이터의 결과

[그림 8]은 D41 데이터를 활용한 시뮬레이션 결과이고, [그림 9]는 D42 데이터를 활용한 시뮬레이션 결과이다. D41과 D42 데이터는 시뮬레이션에 참여한 모든 데이터들 중에 가장 사용정보가 희박하게 나타난 데이터들이다. D41과 D42 데이터를 활용한 시뮬레이션에서 OPBM은 24%의 평균 적중률을 나타내었고, CBFM의 35%의 평균 적중률을 나타내었다. 가장 희소한 선호도를 소유한 D41와 D42 데이터의 문제로 OPBM과 CBFM 모두 낮은 성능을 나타내었다. 본 시뮬레이션 결과를 보더라도 사용정보가 디지털 콘텐츠를 추출하는 데 주요한 요인임을 알 수 있다. D41과 D42 데이터에는 적은 수량의 사용정보가 존재하지만, 모든 시뮬레이션 구간에서 CBFM은 OPBM보다 우수한 성능을 나타내었다. CBFM은 사용자의 성향 예측에 활용하는 사용정보가 더 풍부하기 때문에 확률에 기반한 OPBM보다 우수한 성능을 나타낸 것이다.

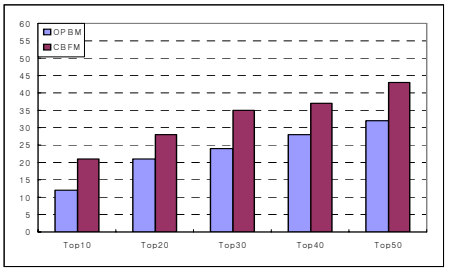

그림 8. D41 데이터의 결과

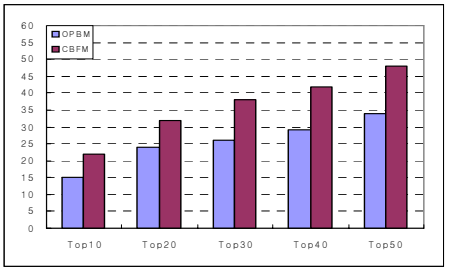

그림 9. D42 데이터의 결과

모든 시뮬레이션 결과에서 CBFM은 OPBM에 비해 우수한 성능을 나타내었으며, 전체 시뮬레이션 결과에서 CBFM의 평균 적중률이 정보량이 가장 많은 D11과 D12에서의 OPBM 평균 적중률과 비슷하게 나타난 결과를 보더라도 CBFM은 희박한 정보량에서도 효과적으로 적용될 수 있다는 것을 알 수 있다.

V. 결론 및 향후 연구

본 논문에서 제안한 소속도를 이용한 사례기반 필터링은 사용정보가 적게 나타난 디지털 콘텐츠 집합에서도 효과적으로 적용할 수 있는 콘텐츠 추출기법이다. 사용자 성향에 적합한 콘텐츠를 추출하기 위해 본 논문에서 제안한 기법은 사용자들의 콘텐츠 사용 사례를 분석하여 콘텐츠의 일반 속성을 확장하고, 속성 확장을 위해 사례기반 소속도를 적용한다.

다른 시뮬레이션 데이터들에 비해 사용정보가 많이 나타난 데이터에서는 CBFM과 OPBM 모두 높은 성능을 나타내었다. 이와 같은 결과가 발생한 이유는 사용정보가 풍부하면 사용자의 성향을 쉽게 파악할 수 있기 때문이다. 사용정보가 적게 나타난 데이터에서 CBFM은 OPBM보다 높은 성능을 나타내었다. CBFM이 OPBM보다 우수한 성능을 나타낸 이유는 OPBM은 발생확률에 기반하여 콘텐츠 추출을 수행하기 때문에 사용자의 성향은 고려하지 않지만, CBFM은 콘텐츠의 사용 사례에 기반하여 사용자의 성향을 고려하고 사용정보가 풍부하기 때문이다. 다양한 시뮬레이션 측정구간에서 CBFM은 OPBM보다 우수한 성능을 나타내었으며, 시뮬레이션을 통해 본 논문에서 제안한 기법이 사용자의 성향을 고려한 디지털 콘텐츠를 효과적으로 추출할 수 있다는 것을 확인하였다.

향후 연구로는 다양한 형태로 존재하는 디지털 콘텐츠를 통합하여 활용할 수 있는 통합 환경에 기반한 필터링에 대한 연구가 필요하다.

참 고 문 헌

[1] 이호영, 정은희, 이장혁, *웹2.0시대 디지털 콘텐츠의 사회적 확산 경로 연구,* 정보통신정책연구원, 2007.

[2] 김진규, 윤재식, 김은정, 정우식, 이강년, 박성만, 김애경, *2010년 2분기 콘텐츠산업 동향분석보고서,* 한국콘텐츠진흥원, 2010.

[3] J. Park and S. Hunting, *XML Topic Maps: creating and using topic maps for the Web*, Addison-Wesley, 2003.

[4] L. Maicher and J. Park, *Charting the Topic Maps Research and Applications Landscape*, Springer, 2006.

[5] J. Park and A. Cheyer, "Just For Me: Topic Maps and Ontologies," In Proceedings of the 1st International Workshop on Topic Maps Research and Applications, pp.145-159, 2005(10).

[6] S. Seedorf, A. Korthaus, and M. Aleksy, "Creating a Topic Map Query Tool for Mobile Devices using J2ME and XML," In Proceedings of the 4th International Symposium on Information and Communication Technologies, Vol.92, pp.111-116, 2005(1).

[7] N. Moraveji, A. Travis, M. Bidinost, and M. Halpern, "Designing an Integrated Review Sheet for an Electronic Textbook," In Proceedings of Conference on Human Factors in Computing Systems, pp.892-893, 2003(4).

[8] R. Wilson, "The Look and Feel of an Ebook: Considerations in Interface Design," In Proceedings of the 2002 ACM Symposium on Applied Computing, pp.530-534, 2002(3).

[9] S. K. Card, L. Hong, J. D. Mackinlay, and E. H. Chi, "3Book: A 3D Electronic Smart Book," In Proceedings of the Working Conference on Advanced Visual Interfaces, pp.303-307, 2004(5).

[10] S. K. Card, L. Hong, J. D. Mackinlay, and E. H. Chi, "3Book: A Scalable 3D Virtual Book," In Proceedings of Conference on Human Factors in Computing Systems, pp.1095-1098, 2004(4).

[11] Y. Chu, I. H. Witten, R. Lobb, and D. Bainbridge, "How to Turn the Page," In Proceedings of the 3rd ACM/IEEE-CS Joint

Conference on Digital Libraries, pp.186-188, 2003(5).

[12] S. V. Agoshkov and P. A. Dmitriev, "Electronic Publication Maintenance Systems," Programming and Computer Software, Vol.28, No.5, pp.293-300, 2002(9).

[13] J. Konstan, B. Millr, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," Communications of the ACM, Vol.40, No.3, pp.77-87, 1997(3).

[14] J. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," In Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence, pp.43-52, 1998(7).

[15] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering," In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.230-237, 1999(8).

[16] D. Billsus, and M. J. Pazzani, "Learning Collaborative Information Filters," In Proceedings of the 15th International Conference on Machine Learning, pp.46-54, 1998(7).

[17] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of Dimensionality Reduction in Recommender System-A Case Study," In Proceedings of the ACM WebKDD Workshop, 2000(8).

[18] M. Deshpande and G. Karypis, "Item-Based Top-N Recommendation Algorithms," ACM Transaction on Information Systems, Vol.22, No.1, pp.143-177, 2004(1).

[19] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," In Proceedings of the 10th International WWW Conference, pp.285-295, 2001(5).

[20] G. Linden, B. Smith, and J. York, "Amazon. com Recommendations: Item-to-Item Collaborative Filtering," IEEE Internet Computing, pp.76-80, 2003(2).

저 자 소 개

김 형 일(Hyungil Kim) 정회원

- 1996년 ~ 1998년 : (주)경기은행
- 2001년 ~ 2004년 : 동국대학교 컴퓨터공학과(공학박사)
- 2005년 ~ 2006년 : 동국대학교 컴퓨터공학과 IT교수(정보통신부)
- 2007년 ~ 현재 : 나사렛대학교 멀티미디어학과 교수

<관심분야> : 추천시스템, 지능형시스템, 인공지능, 데이터마이닝, 임베디드시스템, 기계학습, 의료영상