# *u*-헬스 컨텐츠 추천 시스템

김홍진, 이상철, 김상욱
한양대학교 전자컴퓨터통신공학부
e-mail : wook@agape.hanyang.ac.kr

# Recommender Systems for *u*-Health Contents

Hong-Jin Kim, Sang-Chul Lee, Sang-Wook Kim
Dept. of Electronics and Computer Engineering Hanyang University

**요 약**

*u*-헬스 컨텐츠 증가에 따라 사용자들 자신에게 적합한 정보를 찾기 위해 많은 시간이 요구된다. 본 논문에서는 사용자 프로필 정보와 분야 전문 지식을 바탕으로 사용자에게 가장 적합한 아이템을 추천해주는 시스템을 제안한다. 제안하는 방법을 통해 기존 추천 시스템에서 발생하는 overspecialization 문제와 new user/item 문제를 완화 시킬 수 있다. 실험을 통해 제안하는 추천 시스템이 *u*-헬스 컨텐츠 추천에 적합함을 정성적으로 보인다.

## 1. 서론

최근 건강에 대한 관심이 높아지면서 의학 관련 문서, 미디어 등의 *u*-헬스 컨텐츠의 수가 급격히 증가하고 있다. 그러나 수많은 *u*-헬스 컨텐츠 중에서 사용자에게 적절한 정보를 찾는 것은 쉽지 않다. 따라서 각 사용자에게 적합한 *u*-헬스 컨텐츠를 추천하는 시스템이 필요하다. 추천 시스템(recommender systems)이란 사용자가 아직 평가하지 않은 아이템들 중 사용자의 과거 정보를 사용하여 가장 선호할 것 같은 아이템을 사용자에게 권하는 시스템이다[1]. 추천 시스템은 무수히 많은 아이템들 중에서 사용자에게 개인화된 추천을 함으로써 쉽게 자신이 원하는 것을 선택할 수 있도록 도움을 준다.

본 논문에서는 사용자 프로필 정보와 분야 전문 지식에 기반한 *u*-헬스 컨텐츠 추천 시스템을 제안한다. 우선, 사용자 프로필 정보와 분야 전문 지식을 이용함으로써 기존 추천 시스템의 보편적인 문제점을 완화하는 방법에 대해 설명한다. 그리고 제안하는 추천 시스템에서 추천 결과를 결정하는 알고리즘에 대해 설명한다. 마지막으로, 실험을 통해 제안 시스템이 *u*-헬스 컨텐츠 추천에 적합함을 보인다.

## 2. 관련 연구

추천 시스템은 대표적인 방법으로는 내용 기반 추천 방식과 협업 기반 추천 방식이 있다[1]. 내용 기반 추천 방식은 사용자의 과거 이용 아이템들의 특징과 유사한 아이템들을 추천하는 방식이다. 그러나 사용자가 이용한 아이템의 특징만을 참조해 아이템을 추천하기 때문에 overspecialization 문제가 발생할 수 있다. 또한 과거 이용기록이 없는 새로운 유저에 대해 아이템을 추천하기 어려운 문제가 발생할 수 있다.

협업 기반 추천 방식은 추천 대상자와 과거 아이템 이용 성향이 비슷한 유사 사용자들을 찾고, 이들이 높게 평가한 아이템을 추천하는 방식이다. 그러나 사용자-아이템 간에 평가 기록이 너무 적을 경우 아이템 추천이 어려우며, 새로운 아이템이 생기거나 새로운 사용자가 등록될 경우에도 해당 아이템/사용자에 대한 과거 기록이 없기 때문에 적합한 추천이 어려워진다.

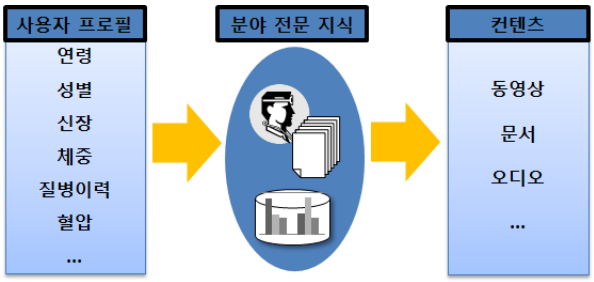
## 3. *u*-헬스 컨텐츠 추천 시스템

### 3.1 제안하는 방법

기존의 내용기반 추천 방식에서는 사용자의 과거 이용 아이템만을 참조해 사용자의 특징을 파악한다. 따라서 사용자가 충분한 수의 이용 기록을 보유해야 사용자의 특징을 정확히 파악 가능하므로 이용 기록이 부족한 새로운 사용자의 경우 특징을 제대로 파악하지 못 하는 문제가 발생한다. 또한, 과거 이용 아이템과 유사한 아이템만을 추천하는 문제가 발생한다.

본 논문에서는 사용자의 이용 기록 대신 사용자 프로필을 이용하여 사용자의 특징을 파악한다. 사용자 프로필은 모든 사용자가 보유하고 있는 보편적인 정보이므로 이를 통해 기존 사용자뿐 아니라 새로운 사용자에게서 다양한 특징들을 도출할 수 있다. 또한, 각 사용자에게 적합한 아이템을 추천하기 위해서 사용자의 특징을 도출해야 한다. 사용자의 특징 도출을 위해 많은 사람에게 인정 받을만한 분야 전문 지식을 이용한다. 신뢰할 수 있는 객관적인 자료를 이용함으로써 추천 결과의 질을 높이고, 사용자의 과거 이용 아이템과 거의 동일한 아이템만을 추천하는 문제도 해결할 수 있다. 분야 전문 지식으로는 질병 통계 자료 및 의사의 진단 기록 등을 이용할 수 있으며, 본 논문에서는 통계청에서 제공하는 국가 통계자료[2]를 이용한다.

그림 1 은 본 추천 시스템의 기본 전략을 보여준다. 타겟 사용자의 프로필을 보고 분야 전문 지식을 참조하여 사용자의 특징을 파악한다. 그리고 사용자 특징에 적합한 *u*-헬스 컨텐츠를 찾고 이를 추천해 준다.




(그림 1) *u*-헬스 컨텐츠 추천 시스템 기본 전략.

제안하는 추천 시스템은 사용자 및 아이템의 특징으로 아이템에서 설명하는 질병 명칭을 사용한다. $u$-헬스 컨텐츠들은 특정 질병을 주제로 제작되었기 때문에 컨텐츠를 대표하는 특징으로 질병 명칭을 사용할 수 있다. 또한, 사용자 프로필 및 분야 전문 지식을 이용하여 사용자의 위험 질병군을 정의할 수 있으므로 사용자의 특징으로도 사용할 수 있다.

식 (1)은 아이템 $s$ 의 특징을 정의한 것이다. $\omega_{si}$는 아이템 $s$ 에서의 질병 명칭 $d_i$의 가중치를 의미하며, 아이템의 메타 데이터 참조 및 TF-IDF[3] 등을 이용하여 계산된다.

$$Content(s) = \overrightarrow{\omega_s} = \{\omega_{s1}, \ldots, \omega_{sn}\} \quad (1)$$

식 (2)는 사용자 $c$ 의 특징을 정의한 것이다. $\omega_{ci}$는 $s$ 에 대한 질병 명칭 $d_i$의 가중치를 의미하며, 분야 전문 지식에 의해 도출된 질병 명칭을 토대로 계산된다.

$$\mathrm{User(c)} = \overrightarrow{\omega_c} = \{\omega_{c1}, \ldots, \omega_{cn}\} \quad (2)$$

식 (3)은 사용자와 아이템 간의 유사도 점수를 구하는 수식이다. 사용자 $c$ 의 $User(c)$와 아이템 $s$ 의 $Content(s)$는 각각 벡터 $\overrightarrow{\omega_s}$와 $\overrightarrow{\omega_s}$로 표현될 수 있다.

$$\mathrm{Score(c,s)} = \cos(\overrightarrow{\omega_c}, \overrightarrow{\omega_s}) = \frac{\overrightarrow{\omega_c} \cdot \overrightarrow{\omega_s}}{\|\overrightarrow{\omega_c}\| \ \|\overrightarrow{\omega_s}\|} \quad (3)$$

사용자 벡터와 컨텐츠 벡터간의 유사도를 계산하는 방법에는 유클리드 거리(Euclidean distance)[4], 코사인 유사도(cosine similarity)[5] 등이 이용될 수 있다. 본 논문에서는 이 중 가장 보편적인 방법인 코사인 유사도로 유사도 점수를 계산한다. 그리고 유사도 점수가 높은 아이템 중 사용자가 원하는 수만큼의 아이템들을 추천해주게 된다.

### 3.2 알고리즘

알고리즘 1 은 $u$-헬스 컨텐츠 추천 시스템의 알고리즘을 나타낸다. (step 1) 가장 먼저, 분야 전문 지식을 통해 추천을 받을 사용자의 프로필로부터 위험 질병 군을 도출해 사용자 벡터 $\overrightarrow{\omega_c}$ 를 만든다. (step 2-4) 추천 가능한 아이템이 $n_t$개일 때, 모든 아이템에 대하여 사용자와의 유사도 점수를 계산한다. (step 5) 그 후, 유사도 점수를 기준으로 아이템을 정렬하여 상위 $k$ 개의 아이템을 선택해 사용자에게 추천한다.

**알고리즘 1.** 추천 시스템의 흐름.

**Input**: 사용자 프로필, 모든 아이템의 벡터 $\overrightarrow{\omega_s}$
**Output** : 유사도가 높은 상위 $k$ 개 컨텐츠
**1**: 분야 전문 지식을 참조해 사용자 프로필로부터 사용자 벡터 $\overrightarrow{\omega_c}$ 생성
**2**: for $j$ = 1: $n_t$
**3**: $\overrightarrow{\omega_c}$와 $\overrightarrow{\omega_{s_j}}$ 간 유사도 점수 $Score(c,s)$ 계산
**4**: end for
**5**: 상위 $k$ 개 아이템 선택

## 4. 실험 방법 및 결과

본 추천 시스템에서는 질병 이력, 신체 정보, 혈압/혈당 측정치를 사용자 프로필로 사용한다. 세 가지 정보를 토대로 사용자 벡터를 생성한다. 사용자가 질병 이력이 있는 경우, 해당 질병이 완치되었다 해도 지속적으로 관리할 필요가 있다. 간암의 경우, 수술 후에도 1 년 이내 재발할 확률이 61%에 달한다[6]. 그 외 심장병, 당뇨병, 고혈압, 고지혈증, 관절염, 골다공증 등도 재발 위험이 높거나 치료가 쉽지 않은 질병이다. 따라서 질병에 대한 이력을 사용자 프로필로 사용한다.

혈압, 혈당의 경우 수치가 가변적이지만, 일정 수치를 넘어가게 될 경우, 주의를 요하게 된다. 혈압의 경우, 수치에 따라 고혈압과 저혈압을 의심해 볼 수 있으며, 혈당의 경우는 당뇨병, 고지혈증을 의심해볼 수 있다. 측정횟수를 늘렸을 때, 계속해서 이상 수치가 나타날수록 해당 질병들에 대한 위험이 높아진다.

정부에서 조사한 국가 통계 자료를 통해 연평균 사망원인, 병원 진료 통계 자료를 구할 수 있다. 해당 데이터를 참고하여 특정 연령, 특정 성별에서 감염 위험이 높은 질병을 파악할 수 있다.

본 논문에서는 실제 서비스되고 있는 $u$-헬스 컨텐츠 620 개를 수집해 실험을 수행하였다. 연령, 성별, 질병 이력, 혈압 수치, 혈당 수치 등 사용자 프로필을 임의로 설정한 가상 사용자 200 명을 대상으로 각 사용자에 대해 $u$-헬스 컨텐츠 620 개 중 5 위까지를 추천하도록 하였다.

표 1 과 표 2 는 사용자에 대한 추천 결과 중 일부를 나타낸 것이다. 50 세 남성에 대해 사망률 1 위인 간암과 간질환 등을 추천하였으며, 혈당에 이상이 있는 45 세 여성에 대해서는 당뇨 및 유방암 등을 추천함으로써 타당한 결과를 보인다. 나머지 사용자에 대해서도 타당한 추천 결과를 보였으며 지면 관계 상 생략한다. 실험 결과, 본 추천 시스템의 추천 결과가 $u$-헬스 컨텐츠 추천에 적합함을 확인하였다.

<표 1> 50 세 남성 추천 결과의 예.

| 50 세 남성, 이력 무, 혈압/혈당 정상 | |
|---|---|
| 순위 | 컨텐츠 |
| 1 | 간암 (영상자료) |
| 2 | 간염과 간암(텍스트) |
| 3 | 간질환 (텍스트) |
| 4 | 위암 (영상자료) |
| 5 | 심장마비(영상자료) |

<표 2> 45 세 여성 추천 결과의 예.

| 45 세 여성, 이력 무, 고혈당 | |
|---|---|
| 순위 | 컨텐츠 |
| 1 | 당뇨병 (텍스트) |
| 2 | 당뇨병 (영상자료) |
| 3 | 유방암 (영상자료) |
| 4 | 고지혈증 (영상자료) |
| 5 | 골다공증 (영상자료) |

## 5. 결론

최근 $u$-헬스 컨텐츠의 수가 급격히 증가하면서 추천 시스템의 필요성이 대두되고 있다. 본 논문에서는 사용자 프로필 정보와 분야 전문 지식에 기반한 $u$-헬스 컨텐츠 추천 시스템을 제안하였다. 그리고 실험을 통하여 실제로 서비스 중인 데이터를 사용하여 본 추천 시스템이 $u$-헬스 컨텐츠 추천에 적합함을 보였다.

## 감사의 글

본 연구는 지식경제부 및 정보통신산업진흥원의 IT 융합 고급인력과정 지원사업의 연구결과로 수행되었음 (NIPA-2011-C6150-1101-0001).

## 참고문헌

[1] G. Adomavicius et al., "Toward the Next Generation of Recommender Systems: A Survey of the State-of the-Art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engineering*, TKDE, Vol.17, No.6, pp.734-749, 2005.
[2] 국가통계포털, 1983-2009.
[3] G. Salton, *Automatic Text Processing,* Addison-Wesley, 1989.
[4] R. Agrawal et al., "Efficient Similarity Search in Sequence Databases," In *Proc. Int'l Conf. on Foundations of Data Organization and Algorithms,* pp.69-84, 1993.
[5] R. Baeza-Yates et al., *Modern Information Retrieval*,.Addison-Wesley, 1999.
[6] 조형원 et al., "간세포암의 절제술 후 재발에 영향을 미치는 인자," 대한외과학회지, Vol.69, No.6, pp.465-570, 2005.