웹 문서의 정보블럭 식별을 통한 효과적인 사용자 프로파일 생성방법

류상현° 이승화 정민철 이은석

성균관대학교 정보통신공학부

{shryu°, shlee, mcjung, eslee}@ece.skku.ac.kr

An Effective User-Profile Generation Method based on Identification of Informative Blocks in Web Document

Sanghyun Ryu° Seunghwa Lee Minchul Jung Eunseok Lee

Sungkyunkwan University Dept. of Information and Communication Engineering

요 약

최근 웹 상에 정보가 폭발적으로 증가함에 따라, 사용자의 취향에 맞는 정보를 선별하여 제공하는 추천 시스템에 대한 연구가 활발히 진행되고 있다. 추천시스템은 사용자의 관심정보를 기술한 사용자 프로파일을 기반으로 동작하기 때문에 정확한 사용자 프로파일의 생성은 매우 중요하다. 사용자의 암시적인 행동정보를 기반으로 취향을 분석하는 대표적인 연구로 사용자가 이용한 웹 문서를 분석하는 방법이 있다. 이는 사용자가 이용하는 웹 문서에 빈번하게 등장하는 단어를 기반으로 사용자의 프로파일을 생성하는 것이다. 그러나 최근 웹 문서는 사용자 취향과 관련 없는 많은 구성요소들(로고, 저작권정보 등)을 포함하고 있다. 따라서 이러한 내용들을 모두 포함하여 웹 문서를 분석한다면 생성되는 프로파일의 정확도는 낮아질 것이다. 따라서 본 논문에서는 사용자 기기에서 사용자의 웹 문서 이용내역을 분석하고, 동일한 사이트로부터 얻어진 문서들에서 반복적으로 등장하는 블록을 제거한 후, 정보블럭을 식별하여 사용자의 관심단어를 추출하는 새로운 프로파일 생성방법을 제안한다. 이를 통해 보다 정확하고 빠른 프로파일 생성이 가능해진다. 본 논문에서는 제안방법의 평가를 위해, 최근 구매활동이 있었던 사용자들이 이용한 웹 문서 데이터를 수집하였으며, TF-IDF방법과 제안방법을 이용하여 사용자 프로파일을 각각 추출하였다. 그리고 생성된 사용자 프로파일과 구매데이터와의 연관성을 비교하였으며, 보다 정확한 프로파일이 추출되는 결과와 프로파일 분석시간이 단축되는 결과를 통해 제안방법의 유효성을 입증하였다.

1. 서 론

최근 정보통신기술의 급속한 발전과 함께 웹 상의 정보는 폭발적으로 증가하고 있으며, 이에 따라 사용자의 요구에 맞는 정보필터링과 개인화서비스가 매우 중요한 이슈가 되고 있다. 특히 전자상거래 분야에서 정보추천 서비스는 사용자의 만족도를 높이고 상점에 대한 충성도를 높이기 위해 필수적으로 요구되고 있다.

전통적인 추천기법으로, 사용자의 관심정보와 컨텐츠의 유사도를 비교하는 컨텐츠 기반 추천방법과 유사한 취향을 가진 사용자들 간에 상호추천을 수행하는 협력적 추천방법이 있다. 이는 모두 사용자의 관심정보를 기술한 사용자 프로파일을 기반으로 한다. 따라서 정확한 사용자 프로파일의 생성은 매우 중요하다.

프로파일을 생성하기 위한 방법은 크게 명시적(Explicit) 방법과 암시적(Implicit)방법으로 분류할 수 있다. 명시적인 방법은 사용자가 상점 방문 초기에 명시적으로 표현한 개인정보나 관심정보를 이용하는 것이며, 암시적인 방법은 사용자의 구매행위나 관심을 보인 행동을 기반으로 사용자의 관심을 추론하는 것이다.

명시적인 방법은 사용자의 관심정보를 빠르게 취득할 수 있다는 장점이 있지만, 사용자를 번거롭게 하며 동적으로 변화하는 사용자의 취향을 반영하기 어렵다는 단점이 있다. 따라서 암시적인 프로파일 생성방법이 보다 바람직하다고 할 수 있다. 그러나 최근 추천시스템의 구조는 각 상점에서 사용자의 행위 정보를 분석하는 형태이며, 따라서 일정기간 사용자가 그 상점을 이용해야만 사용자의 관심정보가 분석 가능하다는 단점을 가지고 있다.

이러한 문제를 해결하기 위해, 사용자 기기에서 사용자의 행동을 관찰하고, 이를 통해 관심정보를 분석하는 방법이 연구되었다[1][2][3]. 대표적인 방법으로는 사용자가 브라우저에 등록한 즐겨찾기 사이트를 분석하거나, 사용자가 이용한 웹 문서를 분석하고 자주 등장하는 키워드를 기반으로 사용자의 관심을 추론하는 연구가 있다.

그러나 최근 웹 문서는 사용자의 취향과 크게 관련이 없는 많은 구성요소들(로고, 저작권정보, 페이지

이동버튼 등)을 포함하고 있다. 이는 동적으로 생성되는 문서에서 더욱 일반적으로 나타난다. 관련연구 [4]에서는 이를 *noncontent blocks* 으로 정의하였다.

따라서 본 논문에서는 사용자의 행동을 지속적이고 통합적으로 분석할 수 있는 사용자의 기기에서, 사용자가 평소에 이용한 웹 문서의 정보블럭을 식별하여 사용자 관심단어를 추출하는 새로운 프로파일 생성방법을 제안한다.

이와 같은 제안시스템의 특징을 통해 사용자의 취향을 보다 정확하게 분석할 수 있으며, 사용자의 정보가 부족한 상점에서도 사용자 기기에서 수집된 취향정보를 참고하여 개인화 전략을 수립하는 것이 가능해진다. 또한 사용자가 평소에 이용하는 웹 문서로부터 관심단어들을 추출함으로써, 동적으로 변화하는 사용자의 취향을 반영한 프로파일 생성이 가능해진다.

본 논문에서는 제안방법의 평가를 위해, 최근 구매활동이 있었던 사용자들이 이용한 웹 문서 데이터를 수집하였으며, TF-IDF방법과 제안방법을 이용하여 사용자 프로파일을 각각 추출하였다. 그리고 생성된 사용자 프로파일과 구매데이터와의 연관성을 비교하였으며, 보다 정확한 프로파일이 추출되는 결과와 프로파일 분석시간이 단축되는 결과를 통해 제안방법의 유효성을 입증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 소개하고, 3장에서는 제안시스템의 구성과 전체적인 동작과정을 설명한다. 4장에서는 실험을 통해 제안방법을 평가하였으며, 결론과 향후 연구를 5장에 기술하였다.

2. 관련연구

사용자의 암시적인 행동정보를 기반으로 취향을 분석하는 대표적인 연구로 사용자가 이용한 웹 문서를 분석하는 방법이 있다. 이는 사용자가 이용하는 웹 문서에 빈번하게 등장하는 단어를 기반으로 사용자의 프로파일을 생성하는 것이다[2][3]. 이를 위해 Term Frequency * Inverse Document Frequency (TF*IDF)기법이 사용된다.

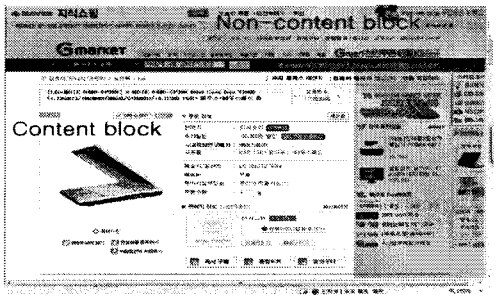
TF는 특정 단어가 문서에 등장하는 빈도를 나타내며, IDF는 전체 문서 집합에 공통적으로 빈번하게 등장하는 단어의 가중치를 낮추기 위해 사용된다. TF*IDF 계산식은 식 (1) 과 같다.

$$W = 1/n\{\Sigma(fik \times [\log(n) - \log(WF) + 1])\} \quad (1)$$

(n = 단어의 수, fik 는 단어의 빈도수, WF는 단어가 나타난 문서 수)

그러나 최근 대부분의 웹 문서들은 사용자의 관심과 크게 관련 없는 많은 부분들을 포함하고 있다. 이는 하나의 사이트를 제작할 때 공통된 레이아웃을 적용하기 때문이며, 특히 동적으로 생성되는 웹 문서의 경우 이 특징은 더욱 일반적으로 나타난다. [그림 1]에 *noncontent blocks* 을 포함하고 있는 웹 문서의 예가 나와있다. 왼쪽 하단에 위치하고 있는 상품정보 부분은 사용자의 취향과 깊은 관련이 있지만, 화면 상단과 오른쪽 부분은 사용자의 관심과 크게 관련이 없는 부분으로 분류할 수 있다.

따라서 웹 문서를 단순히 TF*IDF 방법만을 적용하여 분석하고 사용자 프로파일을 생성할 경우 프로파일의 정확도는 낮아질 것이다.



[그림 1] 웹 문서의 *noncontent blocks* 예

최근 웹 문서에서 이러한 *noncontent blocks* 을 제거하고 정보블럭을 식별하기 위한 연구가 활발히 진행되고 있다[4][5][6]. 이 연구들은 대부분 데스크탑 PC를 고려하여 만들어진 웹 문서를 작은 화면의 휴대용 기기에 적합하게 변환하는 것을 목적으로 하고 있다. 이 중 관련연구 [4]는 웹 페이지의 HTML 태그를 DOM 트리[7]로 변환하고, 각 노드를 순회하며, 반복되는 노드를 식별하여 *noncontent blocks*을 식별한다.

본 논문에서는 프로파일 생성을 보다 효율적으로 수행하기 위해 이와 유사한 접근법을 적용하여, 웹 문서에서 사용자 취향과 관련된 정보블럭을 식별하고, 사용자의 관심단어를 추출한다.

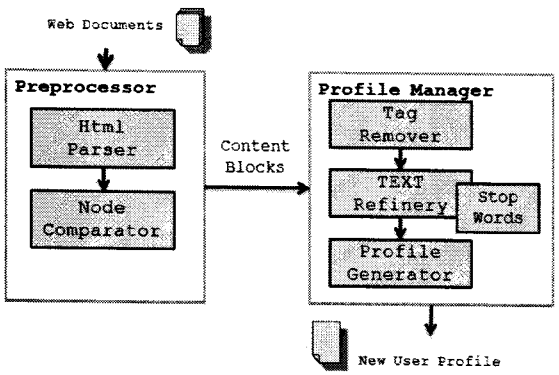
다음 장에서는 TF*IDF방법이나 불용어사전의 구축만으로 한계가 있는 기존 프로파일 생성방법의 단점을 보완하고, 보다 효율적인 사용자 프로파일 생성을 위해 설계된 제안시스템을 자세히 설명한다.

3. 제안시스템

본 논문에서는 사용자가 이용한 웹 문서에서 정보블럭을 식별하여 취향을 분석함으로써, 보다 빠르고 정확하게 사용자 프로파일을 생성하는 것을 목표로 한다.

3.1 제안시스템의 구성

제안시스템은 [그림 2]와 같이 웹 문서를 분석하고 정보블럭을 식별하는 Preprocessor 모듈과, 식별된 정보블럭으로부터 사용자 프로파일을 생성하는 Profile Manager 모듈, 크게 두 부분으로 구성된다. 각 세부모듈의 기능은 다음과 같다.



[그림 2] 제안 시스템의 전체구성

- *Html Parser*: 사용자가 이용한 웹 문서들을 입력 받아 HTML 태그와 텍스트를 분리하고, DOM tree를 구성한다.
- *Node Comparator*: 여러 웹 문서의 URL 앞부분을 비교하여, 동일한 서버로부터 얻어진 웹 문서들을 식별하고, 각 문서의 DOM tree를 따라 노드를 순회하며 일치 여부를 검사한다. 보다 자세한 동작과정은 3.2 에서 소개한다.
- *Tag Remover*: Preprocessor 모듈에 의해 추출된 정보블럭을 입력 받아 HTML 태그를 제거하는 2차 정제 작업을 수행한다.
- *Stop words*: 여러 문서에 빈번하게 등장하지만 사용자의 프로파일에 크게 영향을 주지 않는 단어나 특수문자들의 집합을 의미한다. 영어의 경우 접두사나 접미사, 관사, 대명사 등을 포함하며, 한글에서는 '~하다', '~이다' 와 같은 동사를 포함한다. 또한 웹 문서의 특징분석에 영향을 주지 않는 일반적인 단어들도 불용어로 처리된다.
- *TEXT Refinery*: HTML 태그가 제거된 문자열을 입력 받아 Stop words 리스트를 참고하여 불용어를 제거한다.
- *Profile Generator*: 이전 단계에서 정제된 문자셋을 TF*IDF 기법을 적용하여 처리하며, 기존의 생성된 프로파일에서 가중치가 높은 키워드를 추출하여 새로운 사용자 프로파일을 생성한다. 기존 프로파일에서 추출된 키워드는 새로운 문자셋의 평균 가중치 값이 부여되며, 이를 통해, 기존에 사용자가 선호했던 관심단어들이 일부 보존된다.

3.2 정보블럭의 식별과정

Node Comparator 에 의해 수행되는 정보블럭 식별과정의 알고리즘은 [그림 3]과 같다.

이전 단계에서 *HTML Parser* 는 웹 문서를 HTML 태그와 텍스트를 분리하여 DOM tree를 생성한다. *Node Comparator* 는 각 문서의 URL 정보를 검사하고, 동일한 서버로부터 얻어진 문서 셋을 입력으로 선택한다.

```
Input  : 동일한 서버로 부터 얻어진 문서 셋의
         HTML NodeLists
Output : 정보 블럭
Method :
NodeComparator(NodeList1, NodeList2....)
{
  for(all NodeLists.Node[i])
    if(all node[i] == HTML Tag)
      if(all node[i] same)
        if(nodes.Childnode != Null)
          for(all Child nodes)
            if(!CompareNode(Child nodes))
              return false
            if(all nodes return true)
              return true
        else (not children, nodes same)
          return true
      else(all node[i] not same)
        return false;
    else if(node[i] == Text)
      if(nodes.ChildNode != Null)
        for(all Child nodes)
          if(!CompareNode(Child nodes))
            return false
          if(all nodes return true)
            return true
      else if(nodes.ChildNode == Null)
        return true
      else
        return false
}
```

[그림 3] 정보블럭 식별과정

그리고 각 문서의 노드를 상위노드에서 자식노드로 순회하며, 동일한 내용의 블럭이 있는지 검사한다.

자식노드들의 구조와 내용이 일치한다면, 그 노드를 하나의 *noncontent block* 으로 정의하며, 이를 노드 리스트에서 제거한다.

이러한 과정을 반복함에 따라, 최종적으로는 정보블럭만 남게되며, 결과물은 *Profile Generator* 모듈로 전달된다.

3.3 프로파일의 생성과정

*Tag Remover*는 *Preprocessor* 에 의해 전달받은 정보블럭에서 태그를 모두 제거한다. 이후 불용어 사전을 기반으로 불용어를 제거하고, 남은 단어를 사용자의 관심 단어로 분석한다. *noncontent block* 을 식별하기 위해 불용어 사전을 이용하는 것도 가능하지만, 웹 상의 수많은 문서를 고려하여 불용어 사전을 구축하는 것은 매우 어려운 일이다.

이후, 정제된 정보블럭에 포함된 단어들은 TF*IDF 기법을 적용하여 분석되며, 이를 통해 사용자가 평소에

주로 이용하는 웹 문서의 핵심 키워드를 추출하여 사용자 프로파일을 작성하는 것이 가능해진다.

생성되는 사용자 프로파일은 식 (2)와 같이 정보필터링 분야에서 일반적으로 사용되는 벡터공간 표현방법(Vector space representation)을 이용한다.

$$Profile_A = \{(term_1, w_1), (term_2, w_2) \cdots (term_n, w_n)\} \quad (2)$$

식 (2)에서 w는 각 단어의 가중치를 의미하며, 이는 TF*IDF 방법에 의해 계산된다.

이때 사용자의 기존 프로파일에서 가중치가 높은 키워드를 일부 추출하여, 새로운 문서 셋으로부터 얻어진 키워드들의 평균 가중치를 부여하고, 이를 새로운 프로파일에 포함시킨다. 이러한 특징을 통해 동적으로 변화하는 사용자의 취향을 획득함과 동시에 기존 사용자의 관심단어도 일부 보존 된다.

4. 구현 및 평가

본 논문에서는 제안 시스템의 평가를 위하여, 최근 구매활동이 있었던 사용자가 이용한 웹 문서 데이터를 수집하였으며, TF*IDF 방법만을 이용한 경우(Case1)와 정보블럭 식별과정이 추가된 제안시스템(Case2)을 이용하여 각각 프로파일을 추출하였다. 그리고 추출된 프로파일을 실제 사용자의 구매데이터 페이지와 비교하여 단어의 일치도를 분석하였으며, 정확도 측면과 프로파일분석에 소요되는 시간 이득 측면의 성능평가를 수행하였다.

제안시스템은 윈도우 플랫폼에서 Java를 기반으로 구현되었으며, 사용자가 이용한 웹 데이터는 Temporary Internet Files 폴더에 저장된 웹 문서들을 이용하였다.

4.1 프로파일의 정확도 평가

첫 번째 실험은 제안시스템을 통해 생성된 사용자 프로파일의 유용성 평가를 위해 수행되었다. 이를 위해, 최근 구매활동이 있었던 사용자들이 이용한 웹 문서 데이터를 구매시점에서 2주 전 문서까지 200건을 수집하였다.

그리고 각각의 방식을 통해 사용자 프로파일을 생성하였으며 그 예가 <표 1>에 나타나있다.

<표 1> 생성된 사용자 프로파일 문서의 예

Term	Weight	Term	Weight
고객님	54	삼성	31
3개월	49	무이자	29
A/S	48	네이버	25
구매	43	C2250	24
SCH	41	W2700	23
무료	35	그 외 483개의 용어	

(a) Case 1

Term	Weight	Term	Weight
SCH	32	블루투스	11
C2250	21	LG	8
W2700	20	KTF	10
신규가입	19	구매	9
무료	14	뉴스	8
기기변경	12	그 외 188개의 용어	

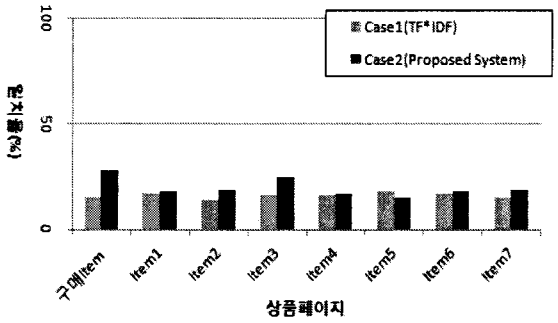
(b) Case 2

<표 1a>는 단순히 TF*IDF 계산방법만을 적용하여 수집한 프로파일의 일부를 표로 나타낸 것이며, <표 1b>는 제안시스템을 통해 *noncontent blocks* 을 제거하고 생성한 프로파일의 일부를 표로 나타낸 것이다. 표에 나타난 것과 같이 정보블럭에서 프로파일을 추출한 경우, 보다 의미 있는 단어들이 프로파일로 추출될 수 있음을 쉽게 확인할 수 있다.

앞에서 언급한 것과 같이 *noncontent blocks* 의 단어들은 불용어 사전에 포함하여 어느 정도 해결할 수 있지만, 이는 매우 많은 인간의 노력을 필요로 하며 한계가 존재한다. 또한 IDF를 사용하여 여러 문서 집합에 보편적으로 나타나는 단어의 가중치를 줄이는 방법도 의미가 낮은 단어의 가중치를 줄이는 데 어느 정도 효과를 기대할 수 있지만, 동일한 사이트에서 반복적으로 나타나는 부분을 제거하는 제안시스템에 비해서 성능이 낮음을 확인하였다.

웹 문서로부터 사용자 프로파일을 생성한 이후, 다수의 상품 페이지에서 사용자 프로파일 생성과 유사한 방식으로 각각 주요 단어를 추출하였다. 이 상품 페이지 집합에는 사용자가 구매한 상품 페이지도 포함시켰다. 그리고 사용자 프로파일과 상품 페이지의 인덱스 파일을 비교하였다.

이 실험결과, 대부분의 사용자 프로파일은 사용자가 구매한 페이지의 주요 키워드를 상당부분 포함하고 있었다. 실험결과를 비교한 도표가 [그림 4]에 나타나있다.



[그림 4] 사용자 프로파일과 상품페이지의 비교

이는 단순히 키워드의 일치도를 비교하였기 때문에, 실험이전에 예상했던 수치보다는 낮은 결과였지만, 실제 추천시스템에 적용하는 경우에 온톨로지(Ontology)가 추가된다면, 본 실험결과보다 훨씬 높은 일치도를 나타낼 것으로 기대된다. 또한 실험결과에서 제안시스템은 TF*IDF만을 이용한 프로파일 생성방법보다 높은 일치도를 나타냄을 확인하였다.

그리고 상품을 구매하기 이전에 그 상품과 관련된 여러 정보를 검색한 사용자의 경우, 높은 수치가 나타났지만, 그렇지 않은 경우에 일치도는 높지 않았다. 또한 실험에는 뉴스 컨텐츠와 같이 상품구매와 크게

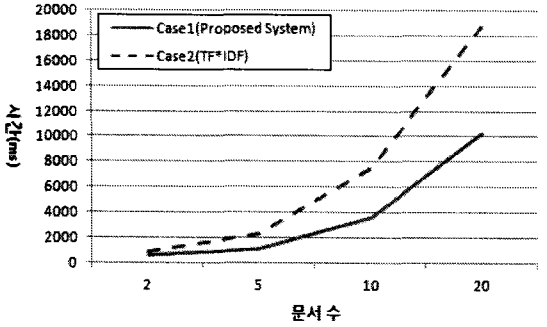
관련이 없는 모든 웹 페이지를 포함하여 프로파일을 추출하였기 때문에 상품 페이지와의 일치도가 낮았다는 것도 확인하였다.

따라서 사용자가 이용한 웹 페이지들을, 그것이 포함하고 있는 단어들을 기반으로 그룹화하고, 여러 주제를 가진 프로파일을 생성하는 경우, 보다 효과적인 추천이 가능해질 것이다.

4.2 프로파일 분석에 소요되는 시간 평가

두 번째 실험은 제안시스템의 프로파일 생성 시간과 TF*IDF의 프로파일 생성 시간을 비교하여 제안시스템의 효율성을 확인하기 위해 수행되었다.

이를 위해, 웹 문서 데이터의 양을 변화시키며 각각 분석에 소요되는 시간을 비교하였다.



[그림 5] 프로파일 생성에 소요되는 시간 비교

본 실험에서는 정확도 실험에서 사용된 사용자의 웹 문서 데이터를 동일하게 사용하였다. 먼저 적은 수의 문서를 이용해 프로파일을 생성하며 소요되는 시간을 체크하였으며, 문서의 수를 점차 증가시키며 시간변화를 체크하였다.

실험결과, 문서의 수가 증가 할수록 프로파일 생성에 제안 시스템의 소요시간은 증가하였다. 그러나 [그림 5]에서 보이는 것과 같이 제안 시스템은 TF*IDF만을 사용한 경우보다 적은 증가율을 나타내었다. 이는 동일한 서버로부터 얻어진 문서의 noncontent block을 프로파일 분석에서 제외시키기 때문이며, 이러한 결과를 통해 noncontent block을 제거하는 것이 프로파일의 생성속도를 단축시킬 수 있음을 확인하였다.

5. 결 론

사용자의 암시적인 행동정보를 기반으로 취향을 분석하는 방법 중 하나로 사용자가 이용한 웹 문서에서 자주 등장하는 단어를 분석하는 기법이 있다. 그러나 최근 웹 문서는 사용자의 취향과 관련 없는 많은 구성요소들을 포함하고 있다. 따라서 이러한 내용들을 모두 포함하여 웹 문서를 분석한다면 생성되는 프로파일의 정확도는 낮아질 것이다.

본 논문에서는 이러한 문제를 해결하기 위해, 동일한 사이트에서 얻어진 웹 문서 집합에 반복적으로 나타나는 부분을 제거하고, 정보블럭을 식별하는 부분을 웹 문서분석 과정에 추가하였으며, 이를 통해 보다 의미 있는 단어들이 사용자의 프로파일로 수집되는 것을 실험을 통해 확인하였다. 또한 정보블럭을 식별하는 과정이 추가되었음에도 불구하고, TF*IDF 방법만을 적용하여 전체 문서를 분석하는 경우에 비해 빠른 처리속도를 나타내는 것을 통해 제안방법의 유효성을 입증하였다.

최근 다수의 가전기기들은 웹을 기반으로 하는 정보기기로 발전하고 있다. 따라서 이 기기들을 사용하는 동안 사용자의 취향을 추론할 수 있는 많은 정보들이 발생하게 되며, 이 정보들이 기기 간에 상호작용을 통해 취합된다면, 제안방법의 적용범위는 무궁무진하다고 할 수 있다.

향후 과제로는 실험결과를 통해 확인된 여러 문제를 보완하기 위해, 웹 문서들을 그 문서가 포함하고 있는 단어를 기반으로 그룹화하고, 다양한 특성을 가진 프로파일을 생성하는 연구를 수행할 것이다. 또한 제안시스템을 통해 생성된 사용자 프로파일을 적용한 효율적인 추천시스템과 텍스트 이외에 멀티미디어 데이터를 포함하고 있는 웹 문서를 분석하여 사용자의 취향을 분석하는 방법에 대해 연구할 것이다.

참고문헌

[1] James Rucker and Marcos J. Polanco, "Siteseer: Personalized Navigation for the Web", Communication of the ACM, vol.40, no.3, pp.73-75, Mar.1997

[2] Thorsten Joachims, Dayne Freitag, and Tom Mitchell, "WebWatcher: A Tour Guide for the World Wide Web", Proceeding of the 1997 IJCAI, Aug.1997

[3] Seunghwa Lee and Eunseok Lee, "A Collective User Preference Management System for U-Commerce", LNCS4773, pp.21-30, Oct.2007

[4] Sandip Debnath, Prasenjit Mitra, Nirmal Pal, and C. Lee Giles, "Automatic Identification of Informative Sections of Web Pages", IEEE Transactions on Knowledge and Data Engineering, vol.17, no.9, pp.1233-1246, Sep.2005

[5] Shian-Hua Lin and Jan-Ming Ho, "Discovering Informative Content Blocks from Web Documents", Proceeding of the eighth ACM SIGKDD Int'l conf. Knowledge Discovery and Data mining, pp.588-593, 2002.

[6] Deng Cai, Shipeng Yu, Ji-Rong Web, and Wei-Ying Ma, "Block Based Web Search," Proceeding of 27th Ann. Int'l ACM SIGIR Conf., pp.456-463, 2004

[7] http://www.w3.org/DOM/