http://dx.doi.org/10.7236/JIWIT.2012.12.3.23
JIWIT 2012-3-4

협업 필터링을 이용한 효율적인 검색 엔진의 설계 및 구현

Design and Implementation of a Efficient Search Engine Using Collaborative Filtering

이기영*, 서일희**, 임명재***, 김규호****, 김정래*****

Ki-Young Lee, Il-Hee Seo, Myung-Jae Lim, Kyu-Ho Kim, Jeong-Lae Kim

요 약 현재 모바일 단말기에 대한 수요가 증가 하고 있으며, 모바일 검색시장이 급속하게 성장하고 있으며 단말기에 따라 적합하게 사용할 수 있도록 모바일 페이지도 등장하고 있다. 하지만 아직까지 모바일 단말기에 최적화된 검색엔진에 대한 연구가 미비한 편이다. 따라서 본 논문에서는 모바일 검색 성능을 향상하기 위해 사용자들의 방문 페이지의 내용, 분야별로 키워드 셋을 구축 후 사용자 성향을 파악하며, 파악된 성향을 통합적으로 관리 및 유사 사용자 정보를 고려해 성향에 맞는 검색 페이지를 추천하였다.

Abstract Recently, due to the increasing demand for mobile devices, mobile searching market is rapidly growing. However, there is the limit of screen size, when searching for mobile devices, various results should be shown at a glance. The reason is that results are important given that up to 43 percent of people tend to check only first page. In this paper, a set of keywords for searching will be used to find out the users' interests. Users were divided into groups after going through Collaboration filtering. Therefore, the result of this experiment, reduced time for searching and improved quality of searching were confirmed.

Key Words : Users' Interests, Collaboration Filtering, Search Engine

Ⅰ. 서 론

최근 모바일 검색에 대한 관심이 높아지고 있는 가운데 다양한 연구가 진행되고 있다[1-2]. 기존에 사용하던 검색이 폭넓고 많은 정보를 제공하였다면, 현재의 모바일 검색은 화면 크기의 제약이나 이동시 사용하는 점을 고려해 정확하고 간단한 정보가 중요시 되고 있다[3-4]. 이러한 모바일 검색 성능을 개선하기 위해 사용자의 성향을 파악해 원하고자 하는 정보를 제공해야 한다[5-6].

따라서 본 논문에서는 사용자의 성향을 파악하기 위해 사용자가 검색한 페이지에 대한 정보를 검색 엔진이 반복적으로 학습하는 전략을 기반으로 사용자에게 최적의 검색 결과를 제공해 줄 수 있는 시스템을 제안하였다. 제안 시스템으로는 사용자가 방문한 페이지에 대한 내용을 분야별로 구축한 키워드 셋과 비교 후 사용자의 성향을 파악한다. 파악된 성향은 통합적으로 관리

*종신회원, 을지대학교 의료IT마케팅학과
**준회원, 을지대학교 의료IT마케팅학과
***종신회원, 을지대학교 의료IT마케팅학과(교신저자)
****정회원, 을지대학교 의료IT마케팅학과
*****종신회원, 을지대학교 의료공학과
접수일자 2012년 5월 7일, 수정완료 2012년 6월 1일
게재확정일자 2012년 6월 8일

Received: 7 May 2012 / Revised: 1 June 2012 /
Accepted: 8 June 2012
***Corresponding Author: lk04@eulji.ac.kr
Dept. of Medical IT and Marketing, Eulji University, Korea

하고, 검색 시 사용자와 유사한 성향을 가진 다른 사용자의 정보를 고려해 최종적으로 사용자의 성향에 맞는 검색 페이지를 추천해 줄 수 있다.

본 논문은 2장에서 관련연구를 기술하며 3장에서는 시스템의 설계 및 구현을 기술하고, 4장에서 실험 및 결과를 기술한다. 끝으로 5장에서 결론을 맺는다.

II. 관련 연구

1. 특징점 검출 알고리즘

웹 에이전트[7]는 관리자의 개입없이 인터넷 상의 정보를 찾거나 이를 분류하는 역할을 하는 프로그램을 말한다. 자동으로 모니터링 된 사용자의 행위를 기반으로 관심 영역을 추출하는 웹 에이전트로 카네기 멜론 대학의 Personal Webwatcher와 앤더슨 컨설팅 랩에서 개발한 InfoFinder가 존재한다. Personal Webwatcher는 사용자의 행위에 따른 키워드 추출을 통해 벡터 테이블을 생성 후 이를 기반으로 TFIDF 및 베이지안 확률(Bayesian Probability)를 적용하여 사용자 프로파일을 구축한다[8]. 다음으로 InfoFinder는 Personal Webwatcher와 비슷하지만 사용자가 직접 관심도를 표현할 수 있는 감독 학습(Supervised Learning)을 사용하고 있다[9].

2. 협업 필터링

협업 필터링은 많은 사용자들로부터 얻은 기호정보(Taste Information)에 따라 사용자들의 관심사들을 자동적으로 예측하게 해주는 방법이다[10]. 협력 필터링 접근법의 근본적인 가정은 사용자들의 과거의 경향이 미래에서도 그대로 유지 될 것이라는 전제에 있다. 협업 필터링은 특정 사용자의 정보에만 국한 된 것이 아니라 많은 사용자들로 부터 수집한 정보를 사용한다는 것이 특징을 가지고 있다. 비슷한 취향을 가진 고객들에게 서로 아직 구매하지 않은 상품들은 교차 추천하거나 분류된 고객의 취향이나 생활 형태에 따라 관련 상품을 추천하는 형태의 서비스를 제공하기 위해 사용된다. 협업 필터링의 사례로는 아마존 닷컴 독자 추천 정보, 인터넷 쇼핑몰의 상품 추천 등이 있다. 정보의 홍수 시대에 협업 필터링 같은 기술은 매우 유용함이 입증 되고 있다. 또한 각각의 카테고리 내 아이템의 수가 방대해 지고 있어, 사용자와 연관성이 있는 아이템을 선택하기 위해 모든 아이템을 일일이 확인한다는 것은 불가능하기 때문에 이러한 상황에서 협업 필터링의 적용 분야는 넓어지고 있는 추세다.

3. 유클리디안 거리 측정 방법

유클리디안 거리는 두 점 사이의 거리를 측정 할 때 사용 되는 방법으로, 이 거리를 이용하여 유클리드 공간을 정의할 수 있다. 이 거리에 대응하는 노름을 유클리드 노름이라 하며, 다음 식 (1)과 같이 정의 한다.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n + q_n)^2}$$

$$\| p - q \| = \sqrt{\| p \|^2 + \| q \|^2 2p \cdot q} \qquad (1)$$

실제 거리를 사용하는 방법 뿐 아니라, 인공지능 분야에서 개체의 속성 사이의 유사도를 구할 때 주로 사용하는 방법 중 하나이다.

유클리디안 거리 측정 방법을 사용한 분류 방법에는 크게 두 가지 방식이 있다. 모든 특징점을 비교하여 가장 짧은 거리에 위치한 점만 분류하는 방법과 임계값이하의 모든 값을 분류하는 방법이 있으며, 본 논문에서는 후자 방법을 사용하여 유클라디안 거리 측정 방법을 사용자 성향별 데이터 셋에 적용해 임계값 이하의 모든 값을 유사 성향으로 분류하여 사용하였다.

III. 시스템 설계 및 구현

1. 시스템 흐름도

본 시스템의 전체적인 흐름도는 그림 1과 같다.

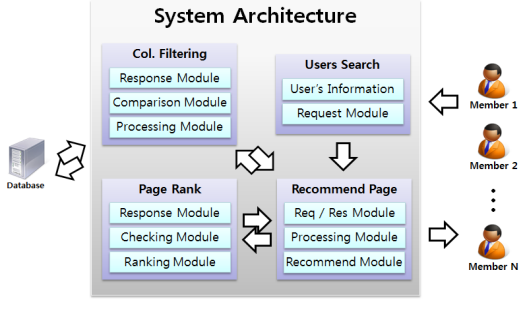

그림 1. 시스템 흐름도
Fig. 1. System Flowchart

먼저 사용자가 입력한 검색 질의를 통해 페이지의 키워드를 추출한다. 그 후 추출한 키워드를 각 분류별 키워드 셋[11]과 매칭 후 반복 학습 해 사용자의 성향을 패턴화 한다. 이렇게 생성된 사용자 성향 패턴을 통해 차후 검색 시 검색의 속도와 정확도를 향상시킬 수 있다.

2. 사용자 성향 파악 알고리즘

본 논문에서는 사용자 성향을 파악하기 위해 사용자 성향별 데이터 셋을 구축 하였으며, 유클라디안 거리측정 방법을 이용하여 유사성향을 측정 하였다. 사용자 성향 간 거리 측정에 사용한 알고리즘은 그림 2와 같다.

```
// Euclidean Space Algorithm
For k=0 to record_num() then
   h=0
   For i=k to record_num()-1 then
      For j=0 to 3 then
         count++
         distance[k][h][j] =
               user_info[k][j] - user_info[i+1][j]

         // 50% 이상 매칭
         if(distance[k][h][j] <=0.9
               0.9 && distance[k][h][j] >= 0.5) then
            distance_group[group_num][0] =
                                 user_info[k][3]
            distance_group[group_num][1] =
                                 user_info[i+1][3]
            group_num++
         End

         // 50% 미만 매칭
         else if(distance[k][h][j] <0.5
               0.4 && distance[k][h][j] > 0) then
            distance_group2[group_num2][0] =
                                 user_info[k][3]
            distance_group2[group_num2][1] =
                                 user_info[i+1][3]
            group_num2++
         End
      Loop
      h++
   Loop
Loop
```

그림 2. 사용자 성향간 거리 측정 알고리즘
Fig. 2. User Tendency Distance Measuring Algorithm

3. 시스템 구현

본 시스템은 사용자가 입력한 검색 질의를 통해 검색된 페이지의 키워드를 추출한다. 이를 각 분류별 키워드 셋과의 매칭을 통해 반복적으로 학습 후 사용자의 성향을 패턴화 한다. 이러한 사용자 성향을 패턴화하고 장기간 학습함으로써 사용자의 성향을 결정하게 된다. 사용자 성향은 검색 엔진 상에 저장되어 다른 사용자가 확인 할 수 없으며, 다른 사용자의 성향을 종합하여 검색 엔진의 질을 향상시킨다.

아래 그림 3은 제안 시스템의 전반적인 구성 모델을 도식적으로 나타낸다.

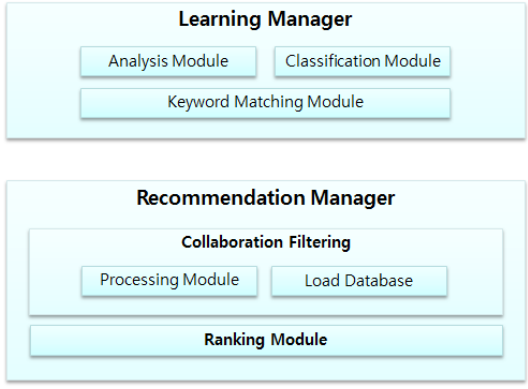

그림 3. 시스템 구조도
Fig. 3. System Architecture

위에서 언급한 제안 시스템의 주요 구성요소는 다음과 같다.

가. Learning Manager

사용자가 입력한 검색 질의를 통해 사용자의 성향을 분류하는 역할을 한다. 하위 모듈로는 분석 모듈, 분류 모듈, 단어 비교 모듈이 존재한다.

나. Recommendation Manager

Learning Manager에서 학습된 결과에 따라 다른 사용자와의 협업 필터링을 통해 검색 결과를 반환하는 역할을 한다. 하위 모듈로는 처리 모듈, 데이터베이스 로드 모듈, 순위 모듈이 존재한다.

Ⅳ. 모의 실험

실험에 사용된 시스템은 표 1과 같으며, 하드웨어 사양은 Intel(R) i5 2.50GHz, 4GB RAM이며, 운영체제는 Windows 7 SP1를 사용하였다. 또한 실험을 위해 오픈 소스 자바 검색엔진인 루씬(Lucene)[12]을 사용하였다.

표 1. 실험 환경
Table 1. Experiment Environment

RAM	4GB
Programming Language	PHP Html5 Javascript
Database	MySQL
Search Engine	Lucene
CPU	Intel i5 2.50GHz

사용자 성향 비교를 위해 사용자 100명에 대한 모의 실험 데이터 셋을 구현하였다. 각각의 사용자는 최대 3가지의 성향을 가질 수 있도록 하였으며, 이러한 성향을 통해 다른 사용자와 성향 비교를 할 수 있도록 하였다. 또한 사용자 성향별 데이터 셋의 일부는 아래 표 2와 같다.

표 2. 사용자 성향별 데이터 셋
Table 2. User Tendency Data Set

학문	인문사회, 공학, 의학, 사회체육, 보건…
컴퓨터,통신	소프트웨어, 하드웨어, 인터넷, 통신, 일반…
엔터테인먼트,예술	연극, 연예 , 건축, 무용, 문학, 사진, 영화…
생활	관광, 레저와 스포츠, 생활, 교통 및 통신…
사회.정치	경영, 경제, 교육, 법, 행정, 정치, 뉴스, 미디어…
스포츠	농구, 야구, 축구, 족구, 배구, 핸드볼, 테니스 …
종교	불교, 개신교, 가톨릭, 힌두교, 이슬람, 기타 종교…
역사	문화재, 지명, 지리, 역사….
철학	논리학, 심리학, 동양철학, 서양철학, 윤리학 …

데이터 셋의 분류 방법은 기존 검색엔진 중 사용자들이 가장 많이 사용하는 구글, 네이버, 다음 등의 분야, 지식 사전 등에서 분류하는 방법을 참조하여 사용하였다.

실험 방법은 사용자들의 각 성향을 1:N으로 비교하여 측정하였으며, 실험 결과는 아래 표 3에서 보는 바와 같이 미일치 비율이 49%로 나타났으나 50% 이상의 매칭률을 보이고 있다.

표 3. 사용자 성향 비교 결과
Table 3. Result of User Tendency Comparison

구 분	50% 이상	50% 미만
1개 일치	1334	966
2개 일치	153	78
3개 일치	3	1
미일치 비율	2515(49%)	

아래 그림 4에서는 그래프로 나타내었고, 1개 일치일 경우는 1,334번의 비교 중 사용자 성향간 50% 이상의 매칭률을 보였고, 2개 일치일 경우는 153번의 비교 중 사용자 성향간 50% 이상의 매칭률을 확인할 수 있었다.

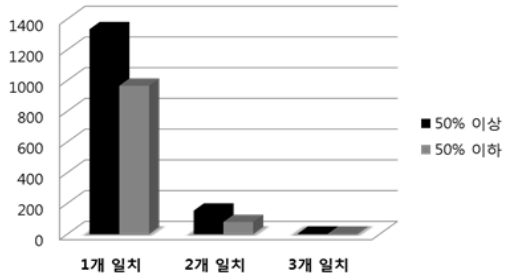

그림 4. 전체 비교 결과
Fig. 4. Total Comparison Results

실험결과, 총 5,050번의 비교 중 사용자 성향간 약 51%(2,535)의 매칭률을 보인 것을 확인할 수 있었으며, 같은 성향을 가진 사용자들에게 정상적으로 웹 페이지를 추천해 주는 것을 확인 할 수 있었다.

Ⅴ. 결 론

현대 사회에서는 개인 특성을 고려한 맞춤형 서비스에 대한 요구가 커지고 있으나, 다양한 영향 요인을 모두 고려하여 적절한 서비스를 결정하기는 어려움이 따른다. 따라서 본 논문에서는 사용자의 성향을 반복적으로 수집하고 분석 및 패턴화하며, 다른 사용자와의 성향 협업필터링을 통해 최종적으로 검색 결과의 질을 향상시키는 시스템을 제안하였다.

사용자 성향 비교 결과 사용자에게 좀 더 정확한 정보를 제공할 수 있는 것으로 파악 되었으나, 미 일치 비

율이 49%정도로 되는 것으로 보아 향후 추가적인 영향 요인을 고려하여 사용자의 성향만이 아닌 다양한 외부 요소들에 대한 연구를 통해 검색 엔진의 성능 향상을 도모하는 방향으로 연구를 진행할 예정이다.

참 고 문 헌

[1] Myung-Jae Lim, Sung-Kyung Hyun, Ji-Eun Park, Ki-Young Lee, "Image Processing for Mobile Information Retrieval Service", The Journal of IWIT, Vol. 11, No. 1, pp. 103-108, 2011.

[2] Soo-Yeoun Kim, Myung-Sin Choi, Chang-Duk Chung, You-Sik Hong, "A Study on the Effects of Mobile Communicational Devices on the Emotional Stability of the Elder Person", The Journal of IWIT, Vol. 9, No. 6, pp. 219-226, 2009.

[3] Ji-Young Lee, Seok-Joo Hong, Hea-Meang Lee, "Implementation and Design the Mobile Client and Server Model(Reverse Really Simple Syndication)", Journal of Korean Institute of Information Technology, Vol. 9, No. 2, pp. 232-244, 2011.

[4] Bo-Hyun Yun, "Dynamic Recommendation of Candidate Attributes for Information Extraction", Journal of Korean Institute of Information Technology, Vol. 9, No. 4, pp. 178-185, 2011.

[5] Seon-Keun Lee, "A Study on the Effective WTLS Processor Design adapted in RFID/USN Environment", Journal of the Korea Academia-Industrial cooperation Society, Vol. 12, No. 6, pp. 2754-2759, 2011.

[6] Hyeog-In Kwon, "A Study on the Software Service Model Evaluation Methodology for Industry Convergence", Journal of the Korea Academia-Industrial cooperation Society, Vol. 12, No. 3, pp.1136-1144, 2011.

[7] Dong-Bum Kim, Byoung-jung Kwak, "Efficient Information Retrieval of A Web Robot Agent on the Internet", Korea Information Science Society Proc. of Autumn, Vol. 29, No. 2, pp. 574-576, 2002.

[8] Dunja Mladenic, "Personal WebWatcher: Implementation and Design", Technical Report IJSDP-7472, 1996.

[9] Bruce Krulwich, "Learning Document Category Description through the Extraction of Semantically Significant Phrases", Center for Strategic Technology Research Andersen Consulting LLP 100 South Wacker Drive, Chicago, IL 60606, 1995.

[10] Michael J. Pazzani, "A Framework for Collaborative, Content-Based and Demographic Filtering", Artificial Intelligence Review 13(5-6): pp. 393-408, 1999.

[11] J. Voss, "Tagging, Folksonomy & Co - Renaissance of Manual Indexing," Proc. of the International Symposium of Information Science, pp. 234-254, 2007.

[12] Otis Gospodnetic, "Lucene In Acation", Acorn Publishing Company, 2005.

저자 소개

이 기 영(종신회원)

• 제 9 권 3호 참조
• 2009년~현재 한국인터넷방송통신학회 이사
• 1991년~현재 을지대학교 의료IT마케팅학과 부교수
<주관심분야 : u-Healthcare, 공간 데이타베이스, GIS, LBS, USN, 텔레매틱스 등>
• Email : kylee@eulji.ac.kr

서 일 희(준회원)

• 2009년~현재 을지대학교 의료IT마케팅학과 학생
<주관심분야 : 인공지능, 바이오 인포메틱스, 검색 엔진, 알고리즘 등>
• Email : nowaday_@naver.com

임 명 재(종신회원) : 교신저자

• 제 9 권 3호 참조
• 1992년~현재 을지대학교 의료IT마케팅학과 교수
<주관심분야 : S/W공학, CBD방법론, HCI 등>
• Email : lk04@eulji.ac.kr

김 규 호(정회원)

• 제 9 권 3호 참조
• 1992년~현재 : 을지대학교 의료IT마케팅학과 교수
• 2011년~현재 을지대학교 산학협력단장
<주관심분야 : u-Healthcare, 유비쿼터스, USN 등>
• Email : khkim@eulji.ac.kr

김 정 래(종신회원)

• 제 10 권 5호 참조
• 현 재 : 을지대학교 보건과학대학 의료공학과 교수
<주관심분야 : 생체정보통신, 생체신호처리 등>
• Email : jlkim@eulji.ac.kr