웹 사용 정보에 기반한 다중 성향 키워드 모델의 설계와 응용☆

Design and Application of Multi Concept Keyword Model based on Web-using Information

윤 태 복* 이 승 훈** 윤 광 호*** 이 지 형****
Taebok Yoon Seunghoon Lee KwangHo Yoon Jee-Hyong Lee

요 약

웹의 방대한 데이터에서 사용자에게 유용한 정보를 제공하기 위하여 다양한 연구가 시도되고 있다. 그중에서 웹 사용 마이닝은 웹 사용자의 로그 정보를 기반으로 의미 있는 패턴을 추출하는 방법이다. 하지만 기존의 웹 사용 마이닝을 이용한 패턴 추출에는 사용자들의 다양한 성향을 고려하지 않은 개별적인 모델을 생성하는데 주를 이루고 있다. 웹에서 사용된 사용자들의 검색 키워드는 그들의 검색 의도나 배경지식에 따라 다양한 의미를 가질 수 있고, 그런 개개인의 검색의도에 맞는 검색 서비스가 제공할 수 있는 기술이 요구된다. 본 논문은 사용자 검색 키워드에 대한 웹 페이지 사용 행위 정보 및 방문한 웹 페이지 리스트를 수집하고 분석하여 다중 성향 키워드 모델(Multi Concept Keyword Model : MCK-Model)을 생성한다. MCK-Model은 사용자들이 특정 키워드를 이용하여 검색 후 방문한 웹 페이지 리스트를 통합하여 생성한 것으로, 사용자들이 검색 키워드에 대해 가지고 있는 다양한 검색 의도에 따라 방문하는 웹 페이지의 정보를 포함하고 있다. 생성된 MCK-Model은 웹 페이지 추천을 위하여 유용하게 사용할 수 있으며, 실험을 통하여 제안하는 방법의 유효함을 확인하였다.

ABSTRACT

There are various studies to provide useful information for users on huge data of web-sites. Web usage mining among them is a method to extract meaningful patterns based on web users' log data. Most of existing patterns of web usage mining, however, had not considered users' diverse inclination but created general models. Web users' keywords can have various meaning upon their tendency and background knowledge. This study is for generating Multi Concept Keyword Model (MCK-Model) by analyzing web usage information on users' keywords of interest. MCK-Model can supply web page network for various inclination based on users' keywords of interest. Also, MCK-Model can be used to recommend the most proper web pages and it has been confirmed that the suggested method is useful enough.

☞ KeyWords : Multi Concept Keyword Model, Web Mining, User Preference, 다중 성향 키워드 모델, 웹 마이닝, 사용자 선호도

1. 서 론

IT기술과 발달과 함께 웹 정보는 기하급수적으로 증가하고 있으며, 대량의 데이터로 부터 사용자는 자신이 원하는 정보를 얻기 위하여 많은 시간과 노력을 들이고 있다. 하지만, 소비하는 시간과 노력에 비해 만족할 만할 결과를 얻기는 쉽지 않으며, 이런 문제를 해결하기 위한 방법으로 패턴 분석, 웹 마이닝 등 다양한 연구가 시도되고

* 정 회 원 : 성균관대학교 컴퓨터공학과 박사과정
tbyoon@skku.edu(주저자)
** 준 회 원 : 성균관대학교 임베디드소프트웨어학과
석사과정 reinblame@skku.edu
*** 준 회 원 : 성균관대학교 컴퓨터공학과 석사과정
yoonkh@skku.edu
**** 정 회 원 : 성균관대학교 컴퓨터공학과 교수
jhlee@ece.skku.ac.kr(교신저자)
[2008/08/18 투고 - 2008/09/09 심사(2009/02/02 2차 - 2009/03/17 3차) - 2009/05/06 심사완료
☆ 이 논문은 2009년도 교육과학기술부의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(No. 2009-0075109)

있다. 웹 환경에서 사용자가 원하는 정보를 보다 지능적으로 서비스하기 위해서 크게 웹 콘텐츠 및 구조를 이해하기 위한 연구와 사용자 웹 사용 정보를 분석하는 방법으로 나뉠 수 있다. 웹 콘텐츠 및 구조이해 방법은 웹 사이트의 이미지, 텍스트, 동영상의 등의 구성 요소를 분석에 활용하거나 웹 페이지의 계층적 구조나 다른 도메인의 웹 페이지간 링크 구조를 분석에 활용하는 방법이다. 사용자의 웹 사용 정보를 활용한 방법은 웹 페이지를 열람하는 사용자의 행위 및 웹 사이트 사용 정보를 분석에 사용한다. 특히 사용자의 웹 사용 정보를 분석하는 연구는 웹 페이지 추천을 위한 기반 기술로 매우 유용하게 사용된다. 예를 들어 사용자들이 방문한 웹페이지를 평가하고, 그 평가 결과를 신뢰도에 반영하여 웹 검색 추천에 사용하는 방법, 또는 사용자가 관심 있게 사용한 키워드나 마우스/키보드 등을 통한 행위 정보를 분석하여 웹 페이지를 선별하고 추천하는 방법 등이 모두 사용자에게 의미 있는 정보를 제공해 주기 위한 연구이다. 하지만 기존의 웹 사용에 따른 평가 및 분석 방법은 다수 사용자의 성향을 고려한 서비스를 제공하기에는 어려운 문제가 있다.

본 논문은 사용자의 관심 키워드 중심의 웹 검색 및 웹 사용 로그 정보를 수집하고 분석하여 다양한 성향 정보를 가지는 키워드 기반의 모델인 다중 성향 키워드 모델(Multi Concept Keyword Model : MCK-Model)을 제안한다. MCK-Model은 사용자 관심 키워드에 대하여 사용자가 방문 했던 의미 있는 웹 페이지들을 이용하여, 사용자들의 다양한 성향 정보를 포함 할 수 있는 연결망이다. MCK-Model의 생성을 위해서 먼저, 사용자 관심 키워드에 기반하여 방문했던 웹 페이지 주소를 수집한다. 수집된 웹 페이지에서 사용자의 행위 정보(마우스/키보드 사용정보, 시간 등)를 이용하여 의미 없는 웹페이지를 제거한다. MCK-Model은 동일한 키워드에 대하여 사용자간에 열람한 웹 페이지 리스트가 유사하다면 결합(Merge)하여 보다 의미있는 연결망을 생성한다. 여기에서 한 사용자의 웹 페이지 열람 정보는 성향이라고 말할 수 있다. 성향(Concept)은 하나의 키워드로 나타낼 수 있는 웹 페이지 연결 리스트를 나타내며, 다중 성향(Multi Concept)은 하나의 키워드에 대하여 여러 가지 성향 정보를 가지는 것을 의미한다. 생성된 MCK-Model은 웹 검색 추천, 키워드 기반 광고, 단어 간 의미 파악 등의 분야에서 유용하게 사용할 수 있는 기술이다.

본 논문의 구성은 다음과 같다. 2장에서는 사용자 웹 검색 추천을 위한 연구 사례에 대하여 소개하고, 3장에서는 제안하는 방법인 다중 성향 키워드 모델의 정의와 생성 방법 및 사용자를 위한 웹 검색 추천에서의 활용에 대해 이야기한다. 4장과 5장에서는 실험과 검증을 통하여 유효성을 확인하고, 끝으로 6장에서는 결론과 향후 연구로 맺는다.

2. 관련 연구

사용자의 웹 사용 정보를 이용하여 웹 사용 행위를 분석하고 추천에 이용하기 위한 연구는 표 1과 같다.

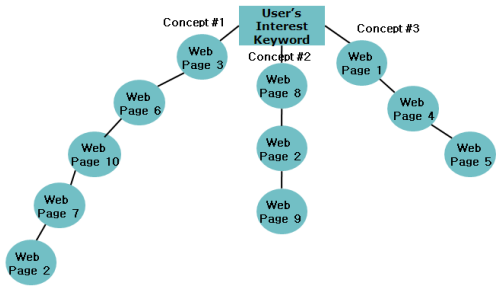
(표 1) 사용자 웹 사용 정보 이용한 연구 사례

연구자	내용
Joh et al.[1], Hay et al.[2]	웹에서 사용자의 활동을 시퀀스로 나타내고 사용자간 유사성을 비교 분석하는 연구
Sufyan, Ahmad[3]	사용자의 행위 정보를 이용한 웹 페이지 평가 방법을 연구
Kang[4]	사용자의 경로 정보 중 필요한 정보만을 찾아 DB를 생성하고 서비스하는 연구
White와 Drucker[5]	단순히 하나의 웹 페이지가 아닌 여러 웹 페이지의 연관된 탐험 행위를 조사 분석구
Eirinaki, Vazirgiannis[6]	사용자의 과거 웹 네비게이션 행위를 기반으로 개인화를 위한 추천 모델을 제시
Chi et al.[7]	협업 필터링과 계층적 k-means 클러스터링 알고리즘을 이용하여 사용자의 웹 사용 습관과 키워드를 분석
이동훈[8]	웹 페이지를 방문한 사용자의 방문 시간과 사용자의 행동 정보를 수집하고, 신경망을 통하여 학습하여 웹 페이지 분류
Motiee et al.[9], Birukov et al.[10]	사용자의 관심 키워드를 이용한 추천 방법을 위해 내용기반 필터링(Content-based Filtering : CBF) 방법을 이용

기존 연구들의 형태는 웹 페이지 사용에 대한 로그 정보를 수집하고 분석하여 패턴을 찾고 웹 사용 정보를 모델링한다. 이 모델은 자동화/지능적/개인화/적응형 등의 서비스를 위한 기반 기술로 활용되지만, 다수 사용자 의 성향이 고려되지 못한 모델 생성으로 사용 범위의 제한적인 모습을 가지고 있다. 다양한 사용자의 성향을 반영한 분석과 모델 생성에 대한 연구의 필요성이 요구된다.

3. MCK-Model의 생성과 활용

본 논문은 사용자의 웹 페이지 사용정보를 수집하고 분석하여 키워드 기반의 웹 페이지 연결망을 생성하는 방법을 소개한다. 하나의 키워드는 다수 사용자의 다양한 성향 정보를 포함하고 있으며, 각 성향 정보에 따라 다른 웹 페이지 연결망을 가지고 있다. 웹 페이지 연결망인 MCK-Model의 정의와 생성 방법 및 활용에 대하여 구체적으로 설명하겠다.

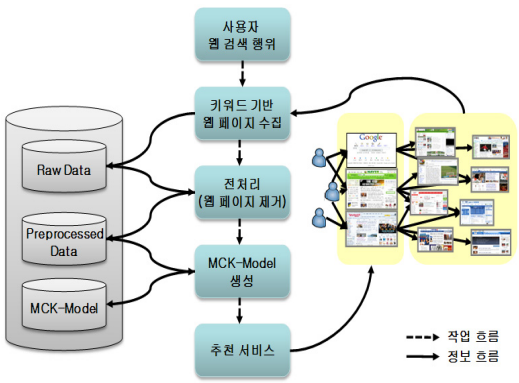


(그림 1) MCK-Model 생성의 예

3.1 MCK-Model이란

MCK-Model은 키워드에 대한 다양한 성향 정보를 포함하고 있는 모델이다. 사용자 개개인이 가지는 성향이나 배경지식의 차이로 어떤 키워드에 대하여 개인에 따라 의미하는 점이 다르다. 앞서 설명했던 “축구”라는 키워드의 예와 같이 사용자에 따라 하나의 키워드가 다양한 성향을 보이게 되는데, 이를 반영한 모델이 MCK-Model이다. 그림 1은 사용자 관심 키워드에 대하여 생성된 MCK-Model의 예이다. 사용자의 관심 키워드에 따라 열람한 웹페이지 10개(Web page 1~10)가 수집되었고, 3개(Concept #1~#3)의 성향으로 분류된 모습을 보이고 있다. “사용자의 관심 키워드”에 따라 Concept #1은 웹 페이지 3, 6, 10, 7, 2의 정보를 가지고 있고, Concept #2는 웹 페이지 8, 2, 9, 그리고 Concept #3은 웹 페이지 1, 4, 5의 정보를 가지고 있다. 만약, 어떤 사용자가 “사용자의 관심 키워드”에 따라 웹페이지 1번과 4번을 방문했다면, 그림 1의 MCK-Model을 이용하여 Concept #3이 선택 되고, 웹 페이지 5번을 추천할 수 있다.

3.2 MCK-Model의 생성 방법



(그림 2) MCK-Model 생성을 위한 전체 흐름도

사용자들의 관심 키워드를 기반으로 웹페이지 정보를 수집하고, 의미 있게 열람한 페이지를 분류하여 사용한다. 사용자는 구글이나 야후 등과 같은 검색 엔진을 이용하여 자신이 원하는 키워드를 입력하고 결과 페이지를 열람하게 된다. 이 때, 사용자의 관심 키워드, 열람한 웹 페이지, 웹 페이지에서의 사용자 행위 등의 정보를 수집한다. 수집된 데이터는 전처리 과정을 거쳐 유효한 웹 페이지를 분류한다. 키워드에 대하여 의미 있는 웹 페이지를 분류하고, 연결망으로 표현한다. 생성된 연결망은 사용자간에 유사도를 측정하여 결합시킨다. 이 결합 작업은 두 가지 장점을 가지고

있다. 첫째, 웹 페이지 연결망의 정보를 간소화 시키며, 저장 관리에 효과적이다. 둘째, 특정 개인의 정보를 다수 사용자와 결합함으로써 생성된 모델의 과적합(Overfitting)을 방지 할 수 있다. 사용자 웹 사용정보를 수집하고, MCK-Model을 생성하기까지의 전체 작업흐름은 그림 2와 같이 나타낼 수 있다. 여기에서 우리는 MCK-Model 생성에 있어서 크게 3가지를 고려해야 한다. 첫째, 사용자가 열람한 웹페이지 중에서 의미 있는 웹페이지를 선별하는 방법, 둘째, 사용자 관심 키워드에 대하여 표현하는 방법, 셋째, 다수 사용자의 웹 페이지 사용 정보를 함축적으로 표현하는 방법이다. 다음 장에서 보다 구체적으로 설명하겠다.

3.2.1 사용자 웹 사용 정보 수집과 의미 있는 웹 페이지 선별

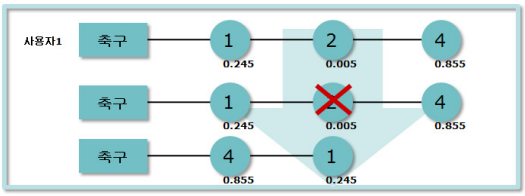
웹 환경에서 사용자들은 자신이 원하는 정보를 얻기 위하여 다양한 검색 엔진(Google, Yahoo, Naver 등)을 통해 웹 페이지에 접근한다. 사용자가 어떤 키워드를 이용하여 검색을 하고 어떤 웹 페이지를 의미 있게 보았다면, 그 정보는 웹 검색 추천을 위한 유용한 정보로 활용 될 수 있다. 사용자 관심 키워드, 사용자 ID, 그리고 사용한 웹페이지에서의 사용자의 행위 정보는 웹페이지가 얼마나 사용자에게 의미 있게 사용하였는지를 측정할 수 있는 요소들이다. 웹페이지를 사용한 사용자에 대해 수집할 수 있는 행위 정보는 사용자를 구분하기 위한 사용자 ID와 관심 키워드를 이용하여 열람한 웹페이지 URL, 웹 페이지 사용 시작 시간, 웹페이지 사용 종료 시간, 다운로드 유무, 복사 & 붙이기 명령 (Ctrl +C, Ctrl +V) 사용 유무, 웹 페이지의 콘텐츠 크기 등 다양하다. 예를 들어 A라는 웹 페이지에서는 3분 동안 머물면서, 다양한 키보드 및 마우스 이벤트가 발생 했고, B라는 웹 페이지에서는 10초간 머물었고 아무런 키보드 및 마우스 이벤트가 발생하지 않았다고 가정하자. 이때 우리는 B라는 페이지 보다는 A라는 페이지가 사용자에게 의미 있는 웹 페이지였다고 생각할 수 있다. 방문한 웹 페이지에서 머문 시간이 얼마 되지 않는다고 하면 사용자가 원하는 내용이 아니라고 판단할 수 있는데, 이런 경우 의미 없는 웹 페이지라고 판단하여 분석에서 제외 시켜야 한다. 또한 웹 로그 수집 과정에서 시스템 오류로 인한 잘 못된 데이터도 마찬가지이다. 웹 페이지가 사용자에게 얼마나 유용했는가를 수치적으로 표현하기 위하여 웹 페이지 점수(Web Page Scoring)[3] 방법을 이용한다. 여기에서 고려해야할 사항으로, 점수 계산에 사용되는 각 요소간의 관계가 얼마만큼 상호간에 영향을 미치는가 하는 것이다. 각 요소는 가중치 값을 이용하여 중요도를 결정한다. 예를 들어 웹 페이지 평가에 사용되는 요소가 웹 페이지 열람 시간, 마우스 클릭, 즐겨찾기 유무 3가지가 있다고 가정하자. 이때 세 가지 요소를 이용하여 웹 페이지가 얼마 유용했는가에 대한 가중치를 얻어야 한다. 동등한 의미를 부여하여 가중치를 계산할 수 도 있겠지만, 경우에 따라서는 시간이 마우스 클릭이나 즐겨찾기의 의미보다 높은 경우도 있고, 또 어떤 경우에는 웹 페이지 사용시간이나 마우스 이벤트보다는 즐겨찾기 행위가 더 중요하다고 여겨 질 때도 있을 것이다. 적용되는 환경에 따라 각 요소의 가중치를 의미 있게 부여하여야 한다. 요소별 가중치 부여를 고려하여 웹 페이지가 사용자에게 얼마나 유용했는가를 측정하기 위하여 아래와 같은 수식을 이용한다.

$$PageWeight_j = 1 - \left(\frac{1}{\sum_{i=0}^{n} (C_i \bullet Attribute_i)} \right) \quad \cdots\cdots\cdots (1)$$

PageWeightj는 사용자가 어떤 키워드를 기반으로 참고한 여러 페이지들 중 j번째 웹 페이지의 가중치를 나타내며, n은 웹 페이지 평가를 위해 사용되는 요소의 개수를 의미한다. Attributei는 i번째 요소를 나타내며, Ci는 i번째 요소의 가중치(상수)이다. 여기서 Attribute은 웹 페이지 사용시간, 마우스 클릭, 마우스 휠 클릭, 마우스 드래그, 키보드 클릭, 복사 횟수 등 웹 페이지에서 수집된

사용자의 행위 정보를 의미 한다.

예를 들어 사용자 1이 “축구”란 키워드를 이용하여 웹페이지 1, 2 그리고 4를 보았다고 가정하자. 이 때 앞서 설명한 PageWeight를 이용하여 웹페이지의 가중치를 계산하였다. 이때 페이지 2의 가중치는 0.005의 값을 가지며, 다른 페이지 보다 아주 현저하게 낮은 수치를 보이고 있기 때문에 제거 되고, 가중치가 높은 페이지 1과 4가 남겨진다.

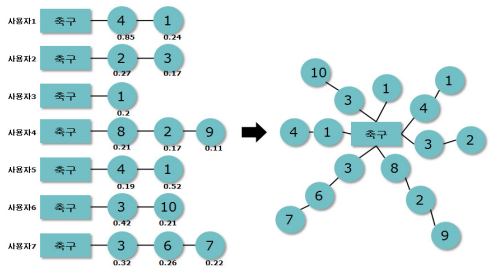


(그림 3) 웹 페이지 가중치를 이용한 의미 없는 페이지 제거

3.2.2 다수 사용자의 성향 정보 표현

그림 3과 동일한 방법을 이용하여 전처리된 사용자 7명에 대한 웹 페이지 집합이 그림 4의 좌측 그림과 같이 수집된 경우, 이는 다시 그림 4의 우측과 같이 통합된 연결망 형태로 표현할 수 있다.

생성된 연결망은 전처리 과정을 거쳐서 의미 없는 웹페이지를 제거하였으나, 사용자의 수가 증가 할수록 연결망의 표현은 복잡하고 거대한 모습을 보이게 된다. 따라서 유사한 웹페이지를 참고한 사용자들 간의 유사성을 이용한 적절한 통합 과정이 필요하다.



(그림 4) (좌) 다수 사용자의 관심 키워드에 대한 웹 페이지 리스트 (우)웹 페이지 리스트의 연결망 표현

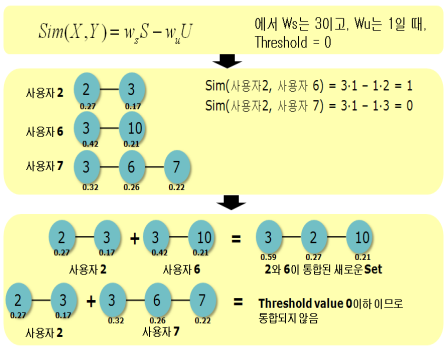
3.2.3 다수 사용자 모델 생성을 위한 병합 (Merging)

만약 n명의 사용자 정보가 수집될 경우 연결망은 n개의 가지(Branch)를 가지게 되며 사용자가 증가할 수 록 연결망의 관리 및 탐색 연산에 드는 비용이 증가하게 된다. 단순히 관심 키워드를 기준으로 사용자가 참고한 웹 페이지의 집합을 나열하는 것을 넘어서 유사한 웹 페이지를 참고한 사용자들 간의 병합이 가능하다면 생성된 연결망을 이해하는데 더 도움이 될 것이다. 그림 4의 경우 7명의 웹 페이지 열람 정보는 7개의 성향으로 표현된다. 연결망의 의미있는 표현을 위해 성향간의 유사도 및 포함관계를 비교하여 통합과정을 거친다. 성향의 병합은 일치형, 포함형, 상호부분일치형으로 크게 3가지 경우로 나뉜다. 먼저 일치형의 경우 두 성향이 동일한 경우를 나타낸다. 그림 4에서 사용자 1과 사용자 5의 경우 웹 페이지 리스트가 동일하다. 이런 경우 한쪽을 제거 한 후 PageWeightj부터 계산된 가중치를 합산한다. 포함형의 경우 사용자 1과 사용자 3에 해당한다. 사용자 1의 정보가 사용자 3의 정보를 포함하는 경우, 사용자 3의 정보를 제거하고 가중치만 합산한다. 마지막으로 상호 부분일치형일 경우 아래와 같은 수식을 이용하여 유사 정도를 판단하고 통합 유무를 결정한다.

$$Sim(X,Y) = w_s S - w_u U \quad \cdots\cdots\cdots\cdots (2)$$

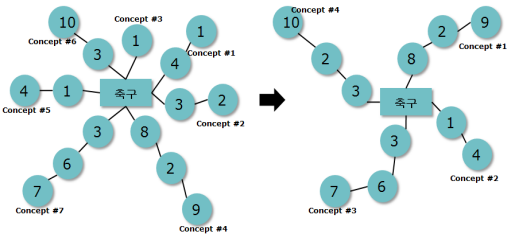
Sim(X,Y)에서 S는 두 Set이 공통으로 포함하는 웹 페이지 개수이고, U는 두 Set이 공통으로 포함하지 않는 웹 페이지 개수이다. 또한, Ws는 두 Set이 공통으로 갖는 웹 페이지에 대한 가중치이고, Wu은 두 Set이 공통으로 갖지 않는 웹 페이지에 대한 가중치를 의미한다. 두 집합의 유사도가 임계값을 넘으면 병합하고, 웹 페이지 가중치는 서로 합하여 하나의 가중치로 만든다. 웹 페이지 유사도 분석을 통한 병합 방법은 그림 5와 같이 중복되는 웹 페이지의 개수와 중복되지 않는 웹페

이지의 개수에 가중치를 각각 곱하여 두 집합의 유사함을 측정하였다. 그림에서와 같이 동일한 경우에 가중치를 3, 틀릴 때 가중치 1이라고 하면, 사용자 2와 6의 경우 중복된 페이지가 1개, 중복되지 않은 페이지가 2개이므로, (3 · 1)-(1 · 2)=1의 유사도를 얻는다. 또한, 사용자 2와 7의 경우 중복되는 페이지가 1개이고, 중복되지 않는 페이지가 3이므로, (3 · 1)-(1 · 3)=0의 유사도를 얻는다.



(그림 5) 사용자 웹 페이지 리스트 간의 유사성 비교를 통한 결합

측정된 두 집합의 유사도는 사용자가 참고한 페이지 리스트를 통합하는 기준으로 사용된다. 만약 임계값(Threshold)을 0이라고 정의 했을 때, 유사도의 값이 0을 초과하면 두 리스트를 통합하고, 0이하이면 통합하지 않는다. 이와 같은 분석 방법을 이용하여 그림 6의 좌측 그림은 그림 6의 우측 그림과 같이 4개의 성향을 나타내는 다중 성향 키워드 모델로 나타낼 수 있다.



(그림 6) (좌) 통합전 웹페이지 키워드 모델 (우) 웹페이지 분석을 통한 MCK-Model의 생성

3.3 웹 검색 추천을 위한 MCK-Model의 활용

그림 6에서 보는 바와 같이 생성된 MCK-Model은 키워드에 기반한 다수 사용자의 성향에 대한 정보를 표현하는 연결망 구조를 가진다. 이 구조는 어떤 키워드에 대하여 하나의 의미만을 가진 웹페이지 정보를 포함하는 것이 아닌, 다양한 사용자들의 의도와 성향에 적절하게 대응할 수 있는 웹 페이지 정보를 포함한다. 어떤 사용자가 "축구"이라는 키워드를 이용하여 웹 페이지를 검색 한다고 가정할 경우 기존의 대다수 방법은 축구라는 키워드에만 의존하여 모든 사용자들에게 동일한 웹페이지를 추천할 것이다. 하지만, MCK-Model은 각 사용자가 가지고 있는 의도에 가까운 검색 결과를 제공해 줄 수 있다. 앞의 예에서 만약 사용자가축구라는 키워드를 이용하여 웹 페이지 3과 6을 방문하였다고 가정한다면, 그림 6 (우) MCK-Model을 이용하여 웹 페이지 7이 사용자가 관심 있는 또 다른 웹페이지라고 할 수 있다. 이는 사용자 검색 키워드에 대하여 의도 및 성향을 고려한 웹 페이지 추천 기술로써 사용자에게 보다 의미 있는 결과를 줄 수 있다.

또한 어떤 키워드가 가지는 의미를 다수 사용자들의 웹 페이지 방문 정보를 통하여 분류하고 모델화함으로써, 키워드를 중심으로 다수 사용자들이 가지는 의도 및 성향에 대한 다양성을 파악할 수 있다. 이는 키워드 연관 검색 및 유사 정보 제공을 위한 기반 기술로 활용 할 수 있다.

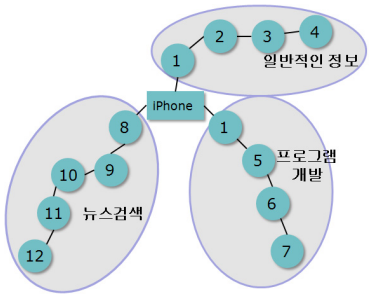
4. MCK-Model 생성 실험

제안하는 방법의 유효함을 확인하기 위해 2가지 실험을 하였다. 첫 번째는 검색에 사용되는 키워드를 사전에 정의하여 사용자에게 제공하고, 검색 정보를 이용하여 분석하였다. 두 번째는 일반적인 웹 환경에서 사용자의 웹 검색 키워드 및 사용 정보를 수집하고 분석하였다. 첫 번째 실험은 제한적인 환경(키워드 및 사용자의 지정)이란 의미에서 "Close-World"라고 하였고, 반대로 두 번째

실험은 제한 조건이 없는 환경에서의 실험이란 의미에서 "Open-World"라고 표현하였다. 사용자들의 로그 데이터 수집을 위해 IE(Internet Explorer)에 Add-on 할 수 있는 프로그램을 제작하였다. 사용자의 웹 사용 정보는 아래와 같다. 사용자의 IP address, IE 버전 7.0 다중 탭 구분을 위한 ID, 열람한 URL, 이전 URL, 웹 페이지 Title, 파일 크기, 문자 길이, 시작시간, 종료시간, 마우스 클릭 횟수, 마우스 더블클릭 횟수, 오른쪽 마우스 클릭 횟수, 마우스 휠 횟수, 드레그 인 드랍 횟수, 키보드 클릭 횟수, 복사 & 붙이기 횟수, 문서 종류, 웹 페이지 프레임별 구분 URL, 클릭된 URL으로 구분되어 수집한다.

4.1 Close-World에서 MCK-Model 생성

실험에서는 구글, 야후, 네이버 검색 엔진의 2006년, 2007년 인기 검색 순위 Top 30 에서 게임 및 특정 사이트 검색을 제외한 키워드 20개 (iPhone, video, 날씨, 대조영, 대출, 된장녀, 디워, 방송사고, 아르바이트, 아찔소, 영화, 월드컵, 중독성게임, 지도, 타자연습, 연예인 N씨 등)를 선별하여 사용하였다. 실험대상은 교내 연구원 중 7명을 선발하여 실시하였다. 수집된 데이터를 보면 전체 823개의 웹 페이지를 방문하였고, 이중 의미 없는 웹페이지를 제거하고 451개 웹페이지를 이용하여 MCK-Model 생성에 사용하였다. MCK-Model을 통하여 141개의 집합을 83개의 집합으로 병합하였다. 그림 7은 MCK-Model을 사용하여 20개 키워드 중에서 "iPhone"의 연결망을 표현한 그림이다. 페이지 1~4를 포함하는 집합은 "iPhone"의 블로그 및 공식 홈페이지와 같은 일반적인 정보를 나타내고 있으며, 웹 페이지 5~7은 "iPhone"의 응용 프로그램 개발과 연관된 커뮤니티 및 개발자 관련 공식 홈페이지이다. 웹 페이지 8~12는 "iPhone"에 관련한 코리안 타임즈 및 IT관련 뉴스 기사에 대한 웹 페이지 정보이다.



(그림 10) "iPhone"의 MCK-Model

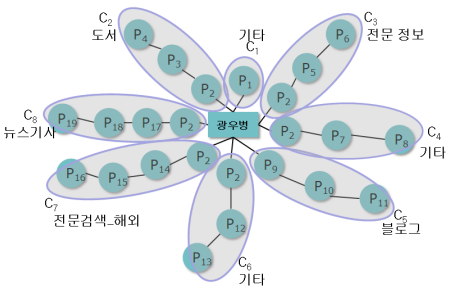
4.2 Open-World에서 MCK-Model 생성

2008년 4월 21일 부터 2008년 5월 23일까지 S 대학교 교내에 자율 PC실 150대에 사용자 웹 사용 정보를 수집할 수 있는 프로그램을 설치하였다. 이 프로그램은 IE(Internet Explorer)를 이용하여 네이버, 야후, 다음, 구글 등의 검색 사이트에서 키워드 입력 후의 웹 페이지 사용 정보에 대하여 수집한다. 전체 수집된 로그 파일 용량은 약 950MByte 이다. 수집된 웹 사용 로그 정보에서 27,965개의 검색 키워드를 찾았고 87,738개의 키워드와 연관된 웹 페이지 사용 정보를 얻을 수 있었다. 27,965개의 키워드 중에서 약 90%이상이 한 번씩만 검색되었던 키워드 이다. 의미있는 MCK-Model 생성을 위해서는 한 개의 키워드에 대하여 다양하게 검색 했던 데이터를 이용하는 것이 유용하게 사용될 수 있다. 그래서, 하나의 키워드에 5회 이상의 검색 정보가 존재하는 경우에만 분석에 이용하였다. 전체 수집정보에서 5회 이상의 검색 정보가 있는 키워드는 395개이다. 대부분의 검색 키워드가 특정 사이트를 찾기 위한 것으로 검색 후 한 번의 클릭으로 웹사이트 이동이 가능한 것이다. 앞서 설명한 바와 같이 어떤 키워드 대해서 모든 사용자가 원하는 절대적인 한 개의 사이트가 존재한다면, 이는 다중 성향이 반영된 MCK-Model 생성에 의미가 없는 데이터이다. 본 실험에서는 국회도서관이나 싸이월드와 같은 특정 기관이나 웹 사이트를 검색하기 위한 키워드를 제외하고 광우병, 전기 분해 등과 같은 사용

자에 따라 원하는 결과가 다양한 키워드를 직접 선별하였다. 분류된 키워드는 113개이며, 이 키워드를 통하여 검색된 횟수는 1,614회, 전체 열람한 웹 페이지의 개수는 3,678개 이다.

113개의 키워드에는 영화, 지도, 자기소개서, 여성과학자, 대운하, 아스피린, 인버터, 광우병, 과외, 광섬유 등 다양하게 나타났으며, 이중에서 '광우병'의 경우 그림 8과 같은 결과를 보였다.

그림 8에서와 같이, 광우병 키워드는 도서, 뉴스기사, 전문 정보, 해외전문정보, 블로그, 기타 분류로 각각의 성향으로 나눠진다. C1, C4, C6의 경우 검색 정보, 블로그, 카페 등 혼합된 웹 페이지 정보를 가지고 있다. C2의 경우 처음에는 검색 페이지 이지만 그 이후 광우병과 관련된 도서 정보 웹 페이지를 참고하였다. C3는 보다 전문적인 정보를 찾는 모습을 볼 수 있었다. C5는 키워드와 관련된 블로그를 방문하였다. C7는 광우병의 최초 발병지인 영국의 환경식품농촌부 정부 기관 사이트에서 정보를 연람하였다. C8은 조선일보 및 YTN에서 키워드와 관련된 뉴스 기사를 열람한 것을 알 수 있었다.



(그림 11) "광우병" 키워드를 이용한 MCK-Model 생성의 예

5. 검 증

제안하는 방법의 검증을 위해 앞에서 생성한 MCK-Model을 이용하여 성능 평가를 실시하였다. 성능 평가 요소에는 MCK-Model의 유효성과 신뢰성 그리고 서비스에 따른 사용자 만족도를 기준으로 하였다. 다음은 3가지 평가 요소에 대한 설명이다.

- 모델의 유효성 : MCK-Model의 컨셉은 연관 있는 페이지들의 집합인가?

사용자 관심 키워드에 따라 열람한 페이지의 집합은 하나의 컨셉으로 표현 될 수 있다. 만약 새로운 사용자가 어떤 키워드에 따른 컨셉이 MCK-Model의 여러 컨셉들 중에 하나와 유사하다면 MCK-Model에 속한 컨셉이 유효하다고 할 수 있다.

- 모델의 신뢰성 : MCK-Model의 페이지들이 의미있는 집합인가?

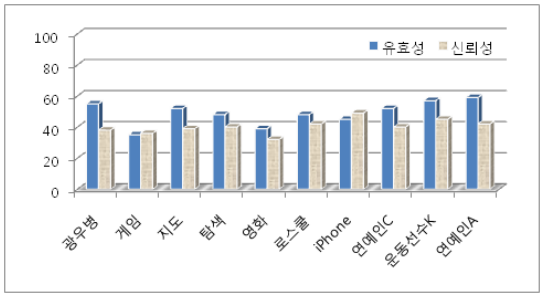
"새로운 사용자의 관심 키워드에 따른 웹페이지 리스트가 MCK-Model에 얼마나 포함되어 있는가?" 와 유사한 질문이다. 웹 검색 엔진(Google, Yahoo, Naver, etc.)을 이용하여 새로운 키워드에 따른 웹 페이지를 열람하였을 때, 이 때 의미 있게 열람한 페이지들의 리스트가 MCK-Model에 많이 포함되어 있다면, 검색 엔진을 이용하는 것 보다 MCK-Model을 이용하는 것이 검색 시간을 절약 할 수 있을 것이다.

- 사용자 만족도 : MCK-Model을 이용한 웹 페이지 추천 서비스에 대한 사용자 만족도가 높은가?

누구에게나 만족스러운 웹 페이지 추천 결과라고 해서, 그 결과가 나에게도 만족스러울 거라는 보장은 없다. 하지만, 다양한 사용자의 성향을 고려한 키워드 모델과 그 모델을 이용한 추천은 서비스는 보다 많은 사용자를 만족 시킬 수 있을 것이다. MCK-Model은 하나의 키워드에 대하여 여러 가지 의미를 가지는 웹 페이지 정보를 가지고 있어, 다양한 사용자의 성향에 맞는 추천을 할 수 있다.

MCK-Model의 검증(유효성, 신뢰성)과 웹 페이지 추천 서비스(만족도)를 위해 MCK Browser를 개발하였다. MCK Browser는 윈도우 기반 응용 프로그램으로 Microsoft Internet Explorer 컴포넌트를 이용하여 제작하였다. MCK Browser는 사용자의

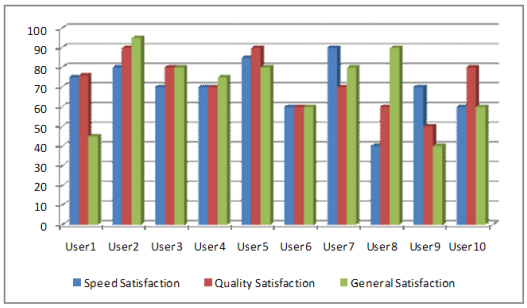
검색 키워드와 열람한 웹 페이지 리스트를 수집한다. 또한, 생성된 MCK-Model을 이용하여 웹페이지 추천 서비스를 제공한다. 웹 추천 서비스를 위해서 사용자의 웹 페이지 열람 정보를 수집하고 MCK-Model의 Pages와 비교하여 현재 사용자의 Concept을 찾고, 아직 열람하지 않은 웹페이지를 추천해준다.



(그림 12) 모델의 유효성과 신뢰성

검증에서는 실험자 10명을 대상으로 일반 Web Browser를 이용한 방법과 MCK Browser를 이용한 방법을 앞에서 언급한 요소들(유효성, 신뢰성, 만족도)에 기준하여 비교하였다. 유효성은 새로운 사용자가 관심 키워드에 따라 열람한 웹페이지 리스트가 MCK-Model에 포함된 Concept의 Pages 리스트와 유사한 정도를 나타낸다. 신뢰성은 새로운 사용자가 열람한 Page들이 MCK-Model에 포함되어 있는 비중을 나타낸다. 유효성과 신뢰성이 높을수록 MCK-Model의 Concept과 Page들이 의미있다고 할 수 있다. 그림 9에서 모델의 유효성은 평균 50이상의 값을 가지며, 신뢰성은 평균 40이상의 값을 가진다. “게임”과 “영화” 키워드가 다른 키워드에 비해 낮게 나온 이유는 키워드가 가지는 의미가 다른 키워드에 비해 너무 추상적이기 때문 다양한 결과를 보이기 때문이다.

MCK Browser를 이용한 웹 페이지 추천 서비스를 경험한 사용자들에게 기존 웹 검색 엔진과의 성능 비교를 설문지를 통하여 조사하였다. 설문지는 원하는 결과를 얼마나 빠르게 얻었는지, 추천 결과에 대해 얼마나 질적(Quality)으로 만족하는지, 그리고 전체적인 만족도를 질문하였고, 0~100 사이의 값으로 표현하게 하였다. 다수의 사용자가 속도, 내용 그리고 전체 만족도에서 평균 70 이상으로 기존 검색 엔진 대비 MCK Browser 서비스를 만족하였다(그림 10).



(그림 13) 설문지를 통한 웹 추천 서비스의 만족도

6. 결론 및 향후 연구

본 논문은 사용자의 웹 검색 키워드에 대한 다양한 성향 정보를 포함할 수 있는 MCK-Model을 생성 방법을 제안하였다. 사용자의 키워드 기반의 웹 사용정보를 기반으로 웹 페이지 연결망을 생성하고, 사용자 성향 간에 유사도를 측정하고 통합하여 보다 의미 있는 연결망을 생성하였다. 생성된 MCK-Model은 웹 페이지 추천서비스에 가능하고, 키워드 중심의 연결망간의 비교/분석을 통하여 의미 유사성을 판단하는 기반기술로 활용가능하다. 또한 실험에서는 지정된 키워드를 이용하고, 특정 대상을 선정하여 웹 검색 행위 정보를 수집하는 방법과 임의 사용자에 대하여 자유로운 키워드 검색을 통한 웹 로그 정보를 수집하는 방법, 두 가지를 경우에 대하여 실험하였다. 두 가지 실험에서 생성된 키워드 기반의 MCK-Model은 사용자 검색 행위에 대한 정보를 잘 나타내고 있었다. 또한, 생성된 MCK-Model의 검증을 위해 사용자를 대상으로 유효성, 신뢰성 그리고 만족도를 평가하였고, 기존 검색엔진 보다 높은 평가를 받았다. 향후 연구로는 MCK-Model 간의 유사도를

측정하여 키워드간의 유사성을 얻어내는 방법과 대량의 사용자 웹 정보에 대한 처리 방법이 요구된다.

참 고 문 헌

[1] Chang H. Joh, Theo A. Arentze, Harry J. P. Timmermans, “A position-sensitive sequence alignment method illustrated for space-time activity-diary data,” Environment and Planning A 2001, vol. 33, pp. 313-338, 2001.

[2] Birgit Hay, Geert Wets, Koen Vanhoof, “Clustering navigation patterns on a website using a Sequence Alignment Method,” Proc. Intelligent Techniques for Web Personalization: 17th Int. Joint Conf. Artificial Intelligence, 2000.

[3] M.M. Sufyan Beg, Nesar Ahmad, “Web search enhancement by mining user actions,” Information Sciences vol. 177, pp. 5203-5218, 2007.

[4] 강귀영, “사용자 경로 정보를 이용한 웹페이지 추천 시스템”, 이화여자대학교 석사학위 논문, 2001.

[5] Ryen W. White, Steven M. Drucker, “Investigating Behavioral Variability in Web Search,” Proc. Int. World Wide Web Conference 2007. 2007.

[6] Magdalini Eirinaki, Michalis Vazirgiannis, ”Usage-based PageRank for Web Personalization,” Proc. 5th IEEE Int. Conf. on Data Mining (ICDM 2005), 2005.

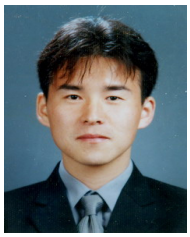
[7] Chen-Chung Chi, Chin-Hwa Kuo, Ming-Yuan Lu, Nai-Lung Tsao, ”Concept-Based Pages Recommendation by Using Cluster Algorithm,” Proc. 8th IEEE Int. Conf. on Advanced Learning Technologies, pp.298-300, 2008.

[8] 이동훈, ”사용자 행동 정보의 수집을 통한 웹 페이지 평가 기법 설계 ”,성균관대학교 석사학위논문, 2009.

[9] Sarah Motiee, Azadeh Nematzadeh,”A Hybrid Ontology Based Approach for Ranking Documents,” Proceedings of World Academy of Science, Engineering and Technology, Vol. 11, Feb. 2006.

[10] Alexander Birukov, Enrico Blanzieri ,Paolo Giorgini,”Implicit: an agent-based recommendation system for web search,” International Conference on Autonomous Agents 2005, pp. 618~624, 2005.

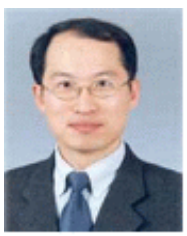
◐ 저 자 소 개 ◑



윤 태 복(Taebok Yoon)
2001년 공주대학교 전자계산학과 졸업(학사)
2005년 성균관대학교 컴퓨터공학과 졸업(석사)
2005～현재 성균관대학교 컴퓨터공학과 박사과정
관심분야 : 사용자 모델링, 게임 인공지능
E-mail : tbyoon@skku.edu

이 승 훈(Seunghoon Lee)
2008년 성균관대학교 컴퓨터공학과 졸업(학사)
2008～현재 성균관대학교 임베디드소프트웨어학과 재학(석사)
관심분야 : 데이터마이닝, 모바일 A.I
E-mail : reinblame@skku.edu

윤 광 호(KwangHo Yoon)
2007년 성결대학교 컴퓨터공학과 졸업(학사)
2008년～현재 성균관대학교 대학원 전자전기컴퓨터공학과 재학(석사)
관심분야 : 웹 마이닝, 기계 학습, 게임 인공지능.
E-mail : yoonkh2000@skku.edu



이 지 형(Jee-Hyong Lee)
1993년 : 한국과학기술원 전산학과(학사)
1995년 : 한국과학기술원 전산학과(석사)
1999년 : 한국과학기술원 전산학과(박사)
2002년～현 재 : 성균관대학교 정보통신공학부 부교수
관심분야 : 지능시스템, 기계학습, 온톨로지
E-mail : jhlee@ece.skku.ac.kr