Proceedings of KIIS Spring Conference 2008 Vol. 18, No. 1.

사용자 웹 사용 정보에 기반한 멀티 컨셉 네트워크의 생성

Multi Concept Network based on User's Web Usage Data

윤광호, 윤태복, 이지형

성균관대학교 정보통신공학부
E-mail: (yoonkh2000, tbyoon,)@skku.edu
jhlee@ece.skku.ac.kr

요 약

웹의 방대한 데이터에서 사용자에게 유용한 정보를 제공하기 위하여 다양한 연구가 시도되고 있다. 웹 사용 마이닝은 웹 사용자의 로그 정보를 기반으로 웹 페이지를 평가할 수 있는 유용한 방법이다. 하지만 기존의 웹 사용 마이닝을 이용한 웹 페이지 평가에는 사용자들의 다양한 성향 패턴을 무시한 일괄적인 모델을 생성하는데 주를 이루고 있다. 본 논문은 사용자 관심 키워드에 대한 웹 페이지 사용 정보를 수집하고 분석하여 멀티 컨셉 네트워크(Multi Concept Network : MC-Net)를 생성한다. MC-Net은 사용자 관심 키워드에 기반한 다양한 성향 정보에 따른 웹 페이지 연결망을 제공한다. 생성된 MC-Net은 웹 페이지 추천을 위하여 유용하게 사용할 수 있으며, 실험을 통하여 제안하는 방법의 유효함을 확인하였다.

Key Words : Multi Concept Network, Web Recommendation, User Modeling

1. 서 론

IT기술과 발달과 함께 웹 정보는 기하급수적으로 증가하는 모습을 보이고 있다. 대량의 데이터로 부터 사용자는 자신이 원하는 정보를 얻기 위하여 많은 시간과 노력을 들이고 있다. 하지만, 소비하는 시간과 노력에 비해 사용자는 만족할 만할 결과를 얻기는 쉽지 않으며, 이런 문제를 해결하기 위하여 다양한 연구가 시도되고 있다. 웹 환경에서 사용자가 원하는 정보를 보다 지능적으로 서비스하기 위해서는 크게 웹 컨텐츠 및 구조를 이해하기 위한 연구와 사용자의 웹 사용 정보를 분석하는 방법으로 나뉠 수 있다. 특히 사용자의 웹 사용 정보를 분석하여 웹 페이지의 유효성을 측정하는 연구는 웹 페이지 추천을 위한 기반 기술로 매우 유용하게 사용된다. 하지만 기존의 웹 사용 마이닝(Web Usage Mining)을 통한 웹페이지 평가 방법은 다수 사용자의 웹 페이지 사용 행위를 분석하여 일괄적이고 획일적인 결과를 생성한다. 다수 사용자의 웹 페이지 사용정보는 다양한 성향 정보를 가지고 있으며, 다양한 성향 정보가 반영될 수 있는 분석 방법이 요구된다. 본 논문은 사용자의 키워드 중심의 웹 검색 및 웹 사용 로그 정보를 수집하고 분석하여 멀티 컨셉 네트워크(Multi Concept Network : MC-Net)를 생성한다. MC-Net은 사용자 관심 키워드에 대한 의미 있는 웹 페이지들의 연결 형태를 사용자들의 성향에 따라 다르게 표현하는 네트워크이며, 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 다양한 연구 사례에 대하여 소개하고, 3장에서는 제안하는 방법인 멀티 컨셉 네트워크(Multi Concept Network : MC-Net)의 정의와 생성 방법 및 사용자를 위한 웹 검색 추천에서의 활용에 대해 이야기한다. 4장에서는 실험을 통하여 유효성을 확인하고, 끝으로 5장에서는 결론과 향후 연구로 맺는다.

2. 사용자 웹 검색 추천을 위한 연구 사례

사용자 관심 키워드에 대하여 적절한 정보 제공을 위한 웹 페이지 추천과 관련된 연구는 아래와 같이 매우 다양한 모습을 보이고 있다. 웹에서 사용자의 활동을 시퀀스로 나타내고 사용자간 유사성을 비교 분석하는 연구[1,2], 사

용자의 웹페이지 사용정보를 분석하기 위하여 사용자의 행위 정보를 이용한 웹 페이지 평가 연구[3], 사용자의 웹페이지 경로 정보를 기반으로 기존 사용자의 경로 정보 중 필요한 정보만을 찾아 DB를 생성하고 서비스하는 연구[4], 단순히 하나의 웹 페이지가 아닌 여러 웹 페이지의 연관된 탐험 행위를 조사 분석하는 연구[5] 등이 실시되었다. 기존의 연구들의 형태는 웹 페이지 사용에 대한 로그 정보를 마이닝하여 패턴을 찾고 웹 사용 정보를 모델링한다. 하지만, 다수 사용자의 다양한 성향이 고려되지 못한 모델 생성으로 제한된 서비스가 제공되는 문제를 가지고 있다.

3. 멀티 컨셉 네트워크의 생성과 활용

본 논문에서 제안하는 멀티 컨셉 네트워크(Multi Concept Network : MC-Net)는 사용자의 웹 페이지 사용정보를 분석하여 키워드 기반의 웹 페이지 연결망을 생성하는 것을 의미한다. 키워드는 다양한 성향 정보를 포함하고 있으며, 각 성향 정보에 따라 다른 웹 페이지 연결망을 가지고 있다. 다음은 MC-Net의 정의와 생성 방법 및 활용에 대하여 설명한다.

3.1 MC-Net 이란?

MC-Net은 키워드에 대한 다양한 성향 정보를 포함하고 있는 네트워크이다. 사용자간에 배경지식이나 가치관의 차이로 각각의 키워드에 대하여 생각하는 점이 사용자 마다 다르다. 예를 들어 "월드컵"이라는 키워드를 이용하여 사용자들이 원하는 정보를 웹에서 검색한다고 가정하자. 어떤 사용자는 월드컵 뉴스 기사를 보기를 희망하는 경우가 있을 것이고, 또 어떤 경우는 월드컵과 관련된 응원 용품이나 축구 용품 등의 구매를 위한 검색이 있을 수 있을 것이다. 이처럼 하나이 키워드에 다양한 성향을 보이게 되는데, 이를 반영한 모델이 MC-Net이다.

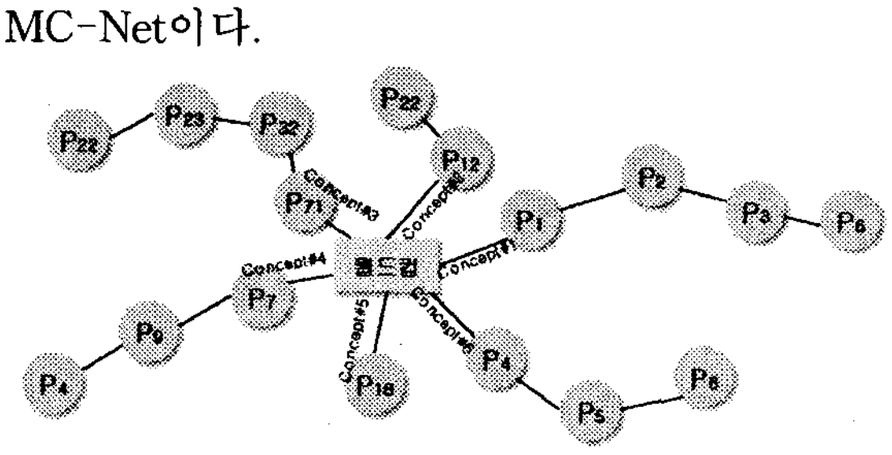

그림 1. MC-Net의 생성 예

다수 사용자의 웹 사용 정보를 기반으로 키워드 성향 네트워크를 나타내는데, 그림 1과 같이 표현한다. P_1 ~ P_n은 "월드컵"이라는 키워드를 이용하여 사용자들이 의미 있게 이용한 웹 페이지들이다. 그림에서는 $Concept_1$ ~ $Concept_6$ 이 생성되었으며, 이는 P_1 ~ P_n의 분석을 통하여 얻어진 6가지의 성향 정보를 나타낸다.

3.2 MC-Net의 생성 방법

앞서 설명한 바와 같이 MC-Net는 사용자들이 이용한 키워드를 기반으로 웹페이지 정보를 수집하고 의미 있게 본 페이지를 분류하여 분석에 사용한다. 사용자 웹 사용정보를 수집하고, MC-Net 생성하기 까지 그림 2와 같이 나타낼 수 있다.

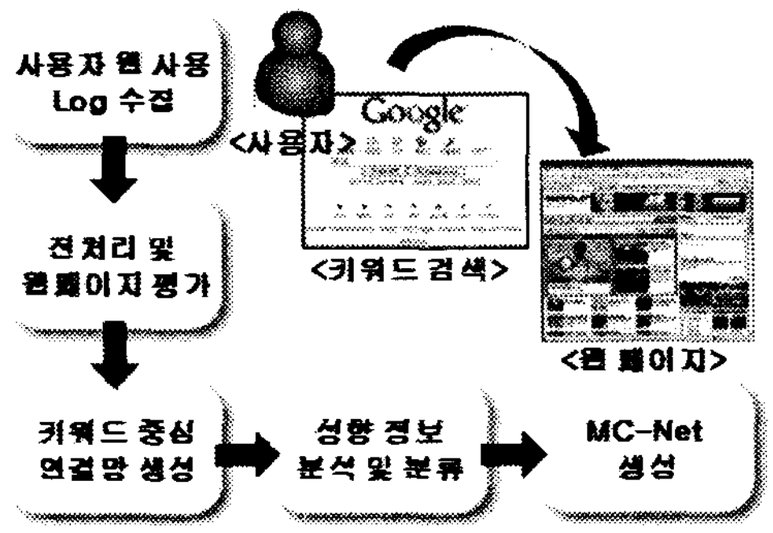

그림 2. MC-Net 생성을 위한 작업 흐름도

- 사용자 웹 사용 Log 수집

웹 환경에서 사용자들은 자신이 원하는 정보를 얻기 위하여 다양한 검색 엔진(Google, Yahoo, Naver 등)을 이용하여 웹 페이지에 접근한다. 사용자가 어떤 키워드를 이용하여 검색을 하고 특정 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 유용하게 사용되었는지를 측정할 수 있는 요소들이다. 웹페이지를 사용한 사용자의 수집할 수 있는 행위 정보로는 사용자 ID와 관심 키워드를 기준으로 사용한 웹페이지 URL, 페이지 사용 시작 시간, 웹페이지 사용 종료 시간, 다운로드 유무, Copy & Paste 명령 (Ctrl +C) 유무, 즐겨찾기 추가 유무, 웹 페이지의 컨텐츠 크기 등 다양하다.

- 전처리 및 웹페이지 평가

사용자의 관심 키워드에 따른 수집된 웹페이지 사용 로그 정보를 이용한 분석에 앞서, 전처리(Preprocess)작업이 필요하다. 사용한 웹페이지의 시간이 너무 작다고 하면 사용자가 원하는 내용이 아니라고 판단할 수 있는데, 이런 경우 분석에서 제외 시켜야 한다. 또한 웹 로그 수집 과정에서 시스템 오류로 인한 잘 못된

데이터도 분석에서 제외시켜야 한다.

웹 페이지가 사용자에게 얼마나 유용했는가에 대한 정량적 표현을 위하여 웹 페이지 점수(Web Page Scoring)[3] 방법을 이용한다. 여기에서 중요한 것은 점수 산정에 사용되는 각 요소간의 관계가 얼마만큼 상호간에 영향을 미치는가 하는 것이다. 일반적으로 점수는 0~1의 값으로 결정하는데, 각 요소는 가중치 값을 이용하여 중요도를 결정한다. 본 논문에서는 각 요소의 의미를 동등하게 보고 가중치를 부여하였다. 페이지가 사용자에게 얼마나 유용했는가를 측정하기 위하여 아래와 같은 수식을 이용하였다.

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

$PageWeight_j$는 사용자가 어떤 키워드를 기반으로 참고한 여러 페이지들 중 j번째 웹 페이지를 나타내며, n은 웹 페이지 평가를 위해 사용되는 요소(시간, 즐겨찾기 유무, 등 사용자 웹 행위)의 개수를 의미한다. $Attribute_i$는 i번째 요소를 말하며, C_i는 i번째 요소의 가중치(상수)이다.

$PageWeight_j$ 는 0에서 1사이의 값을 가지며, 1에 가까울수록 사용자가 의미 있게 본 웹 페이지라고 할 수 있다.

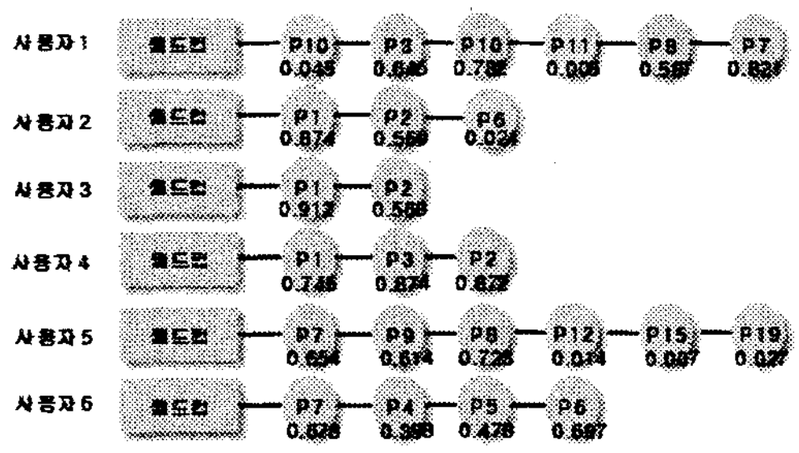

그림 3. 사용자별 웹페이지 사용에 따른 가중치 부여

예를 들어 "월드컵"이란 키워드를 이용하여 6명의 사용자로부터 그림 3과 같은 결과를 얻었다고 가정하자. 여기서 붉은 색으로 표기된 페이지는 점수가 낮기 때문에 분석에서 제외된다.

- 키워드 중심 연결망 생성

그림 3의 사용자별 키워드에 대한 웹페이지 집합은 그림 4와 같이 통합된 키워드 네트워크로 표현할 수 있다.

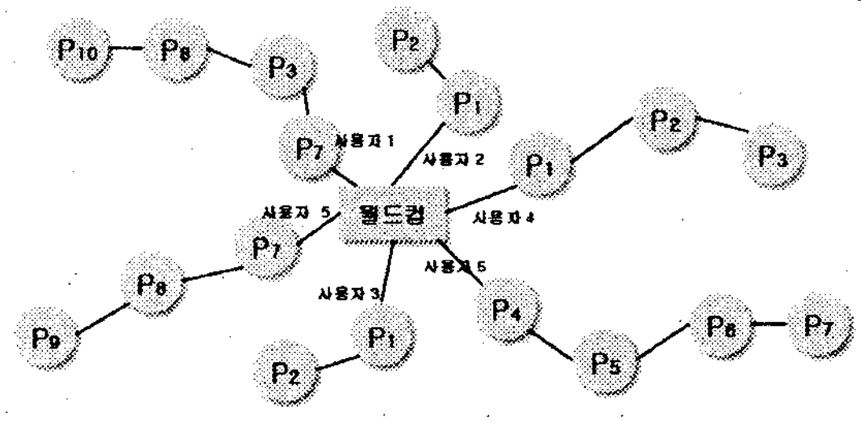

그림 4. 다수 사용자 웹 사용 네트워크

생성된 네트워크의 경우, 전처리 과정을 거쳐 의미 없는 웹페이지는 제거되었으나, 사용자 개개인에 따른 연결망이 생성되어 복잡하고 거대한 모습을 보이게 된다. 분석을 통하여 유사한 웹페이지를 참고한 사용자들 간의 통폐합 과정을 거친다.

- 성향 정보 분석 및 분류

관심 키워드를 기준으로 단순히 사용자가 참고한 웹 페이지의 집합을 나열하는 것을 넘어서 유사한 웹 페이지를 참고한 사용자들 간의 함축적인 표현이 가능하다면 생성된 네트워크를 이해하는데 더 도움이 될 것이다.

네트워크를 요약정리 하기 위해서는 먼저 두 사용자 집합을 선별하고 두 집합을 비교한다. 그림 5에서 사용자 2와 사용자 4의 경우를 예를 들어 설명하겠다. 사용자 2는 웹 페이지 P_1과 P_2를 이용하였고, 사용자 4는 P_1, P_2 그리고 P_3 웹 페이지를 이용하였다.

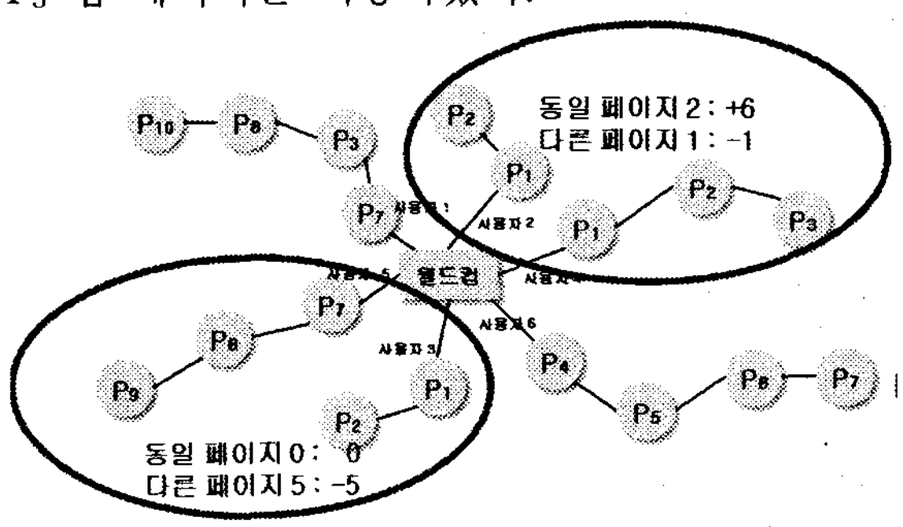

그림 5. 웹 페이지 집합의 유사성 비교

그림 5의 우측 상단 붉은 원에서와 같이 P_1과 P_2 두개가 동일하고 P_3는 같지 않을 때 중복되는 웹 페이지의 개수와 중복되지 않는 웹 페이지의 개수에 가중치를 곱하여 두 집합의 유사함을 측정하였다. 예를 들어 동일한 경우 가중치 3, 틀릴 때 가중치 1이라고 하면 (2*3) + (1 * (-1)) = 5이다. 또는, 사용자 3과 5의 경우 중복되는 페이지가 0개이고, 중복되지 않는 페이지가 5이므로, (0*3) + (5*(-1)) = -5이다. 측정된 두 집합의 유사도는 사용자가 참고한 페이지 집합을 합치는 기준으로 사용된

다. 그림 5의 웹 페이지 연결망을 분석하여 그림 6과 같이 "월드컵"이라는 키워드에 3개의 성향을 나타내는 멀티 컨셉 네트워크(Multi Concept Network : MC-Net)가 생성되었다.

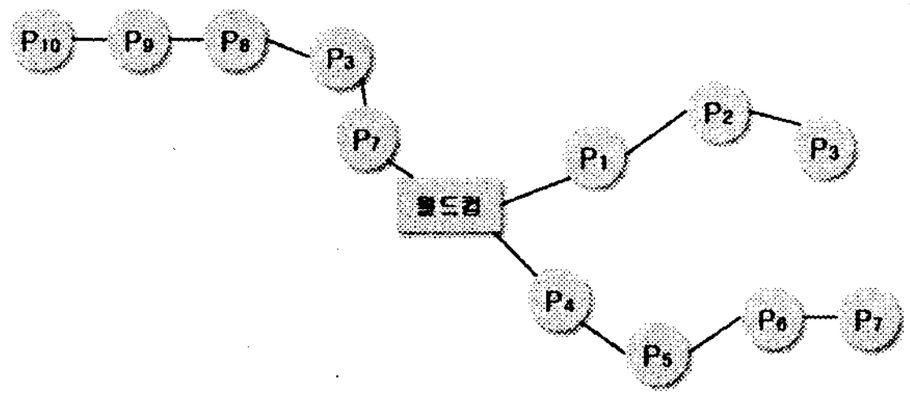

그림 6. 멀티 컨셉 네트워크의 생성

3.3 웹 검색 추천을 위한 MC-Net의 활용

그림 6에서 보는 바와 같이 생성된 MC-Net은 키워드에 기반하여 한 가지 성향에 대한 연관 웹 페이지 정보만을 가지는 것이 아닌 다양한 성향에 대한 정보를 표현하는 네트워크 구조를 가진다. 어떤 키워드에 대하여 하나의 의미만을 가진 웹페이지 선별이 아닌, 사용자가 의도한 성향에 적절하게 대응할 수 있는 정보를 포함한다. 만약 사용자가 "월드컵"이라는 키워드를 이용하여 참고한 페이지가 P_3과 P_7이라고 하면, 웹 페이지 P_8이나 P_9 또는 P_{10}을 추천할 수 있을 것이다.

4. 실험

실험에서는 구글, 야후, 네이버 검색 엔진의 2006년, 2007년 인기 검색 순위 Top 30 에서 게임 및 특정 사이트 검색을 제외한 키워드 20개를(표 1) 선별하여 사용하였다. 특정사이트(로또, 국세청, EBS 등)를 접속하기 위한 키워드나 게임(서든어택, 던전앤파이터 등)플레이를 목적으로 하여 사용한 키워드의 경우 검색 결과에 대하여 한 번 클릭(One-Click)으로 사용자가 원하는 사이트로 이동하게 된다. 어떤 키워드 대해서 모든 사용자가 원하는 절대적인 한 개의 사이트가 존재한다면, 추천의 의미가 없다고 할 수 있다. 실험대상은 교내 연구원중 7명을 선발하여 실시하였다. 수집된 데이터를 보면 전체 823개의 웹 페이지를 방문하였고, 이중 의미 없는 웹페이지를 제거하고 451개 웹 페이지를 이용하여 MC-Net 생성에 사용하였다.

표 1. 실험에 사용된 20개의 키워드

iphone	지도	된장녀
Video	대출	방송사고
디워	아르바이트	아찔소
정일우	영화	슈퍼주니어
원더걸스	날씨	노현정
월드컵	타자연습	이준기
대조영	중독성게임	

MC-Net을 통하여 141개의 집합을 83개의 집합으로 결합하였다. 그림 8은 MC-Net을 사용하여 키워드 '노현정'의 네트워크를 표현한 그림이다.

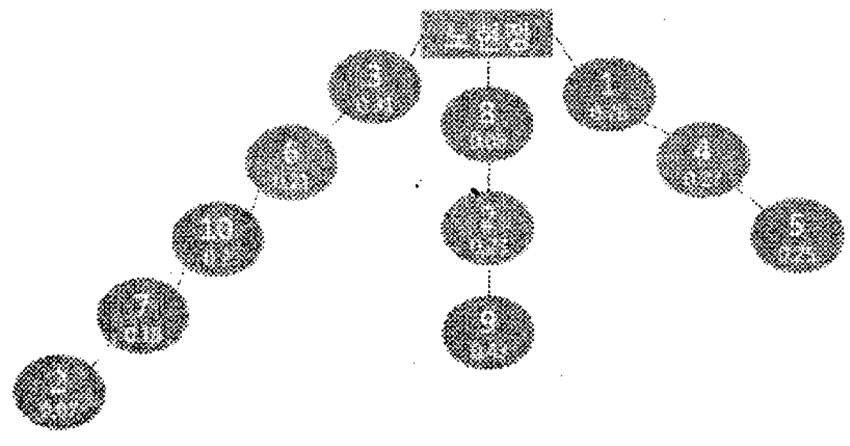

그림 8. 키워드 '노현정'의 MC-Net

페이지 1, 4, 5를 포함하는 집합은 노현정의 임신과 이혼에 관한 기사였으며 페이지 8, 2, 9는 노현정의 결혼 전 기사, 페이지 3, 6, 10, 7, 2는 노현정에 대한 포괄적인 기사였다.

5. 결론 및 향후 연구

본 논문은 키워드에 대한 다양한 성향 정보를 포함하고 있는 네트워크 MC-Net을 제안하였다. 또한 사용자의 검색 행위 분석을 통하여 키워드 별로 MC-Net을 생성하는 것이 가능함을 보였다. 생성된 네트워크는 광고, 웹 페이지 추천, 키워드 의미 분석을 위한 기반 기술로 활용이 가능하다.

향후 연구로는 제한적인 환경(키워드, 사용자 등)이 아닌 일상생활의 사용자 웹 사용정보를 장기간 수집하고 분석하는 지금보다 확장된 형태의 실험이 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, pages 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," The International World Wide Web Conference 2007.