개인화된 토픽 맵의 확장을 통한 연관 토픽의 효과적인 순위화 방법

강형구○ 윤성웅 박정우 이상훈
국방대학교
marine4031@korea.com, ysw1209@gmail.com, jw2236@dreamwiz.com, hoony@kndu.ac.kr

Effective Topic Ranking Method using Extention of Personalized Topic Map

Hyoung-Goo Kang○ Soung-Woong Yoon, Jung-Woo Park, Sang-Hoon Lee
Korea National Defense University

1. 서 론

웹 정보의 폭발적인 증가는 사용자가 필요로 하는 정보를 찾기가 어려운데이러한 정보를 빠르면서도 정확하게 찾아내고자 하는 연구가 활발히 진행중이다.웹 정보의 주 제공원인 포탈(Portal)들은 사용자들이 많이 찾는 검색어(query)의 순위를 별도의 인터페이스(interface)를 이용하여 제공하고 있으나 이는 개별 사용자의 검색 의도를 모두 충족시키지는 못한다. 즉 검색 엔진은 사용자 개개인의 관심(Interest)과 검색상의 문맥(Context)을 수집된 자료를 바탕으로 제한적으로만 고려하고 있다.

이러한 문제점을 해결하기 위해 정보의 자동적인 의미 분석을 가능하게 하는 시멘틱 웹(Semantic Web)이 등장하였는데, 이를 웹 정보에 구현한 대표적 예로 토픽 맵(Topic Map)이 있다.[1] 토픽 맵은 지식과 정보를 정의하고 표현하는 한 방법으로서 분산 환경 하에서 지식 구조를 토픽간의 관계(Association)로 정의하고, 정의된 구조와 자원을 연결(Occurrence)하여 표현한다. 이 구조는 기존 정보의 형태를 변환하지 않고도 토픽간의 관계를 통하여 서로 다른 토픽의 병합이 가능하고 관계 기능을 확장하면 무한에 가까운 정보의 제공과 광범위한 연결이 가능하다는 장점이 있다. 하지만 토픽 맵은 토픽 간 또는 자원과의 관계만을 표시할 뿐 관계의 강도(strength)는 표시하지 않는다. 또한 웹 검색과 마찬가지로 개별 사용자의 선호도를 반영한 정보는 제공하지 않고 있다.[2]

본 논문에서는 개인화된 토픽 선호도를 이용하여 사용자가 선택한 토픽과 연관되는 토픽의 순위화를 연구하였다. BM 적용 시 선행 연구[1]를 확장하여 선택된 토픽과 직접 관계가 있는(1차 관계) 토픽과 함께 직접 관계된 토픽과 관계된 (2차 관계) 토픽까지 고려하였고, 관심 토픽에 관계된 토픽들의 가중치를 결정하기 위한 방법으로 그래프 구조를 트리 구조로 재편성하는 방법을 사용하였다.이는 PTR 알고리즘에서 관심 토픽을 기준으로 연결되는 토픽의 수가 어떤 의미를 가지는지 밝혀내고 확장된 노드를 반영하는 E-PTR 알고리즘을 제안한다.

2. 토픽의 순위화

2.1 PTR 알고리즘의 확장 (E-PTR)

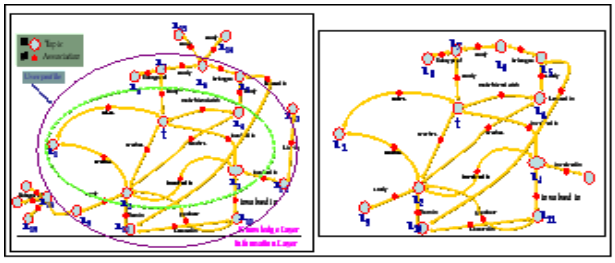
PTR (Personalized Topicmap Ranking) 시스템[1]은 사용자가 토픽에 대한 선택 행위(Click event)를 할 때마다 실시간으로 사용자 프로파일 데이터베이스에 클릭 횟수 정보를 바탕으로 T_i가 계산되어 저장되고, 동시에 가중치(W)에 반영하여 토픽 맵에서 선택한 토픽과 관련된 토픽들을 순위화 하였다.[3][4][5][6][7][8]

$$W_{t,i}^{1} = Ti[\log\frac{(u_i+0.5)(N-n_i+0.5)}{(n_i+0.5)(U-u_i+0.5)}] \quad (1)$$

where, $W_{t,i}$: 관심 토픽 t에 대한 i번째 토픽의 가중치
T_i: i번째 토픽에 대한 사용자 선호도 벡터
N: 토픽 맵 전체의 토픽 수
n_i: 토픽 맵 전체에서 i번째 토픽과 관계되는 토픽 수
U: 사용자 프로파일 내의 전체 토픽 수
u_i: 사용자 프로파일 내에서 i번째 토픽과 관계 되는 토픽 수

<그림 1>은 관심 토픽 't'와 연관된 토픽들의 관계를 도시한 것이며 사용자 프로파일이 PTR (Personalized Topicmap Ranking) 시스템[1]에 따라 실선 내부와 같이 형성되었다고 가정할 시, 토픽"t"와 관련되는 각 토픽의 가중치는 수식 (1)을 활용하여 각 요소의 값을 통해 구할 수 있으며 랭킹 결과는 <표 1>과 같다.

<그림 1> 사용자 프로파일 (예)

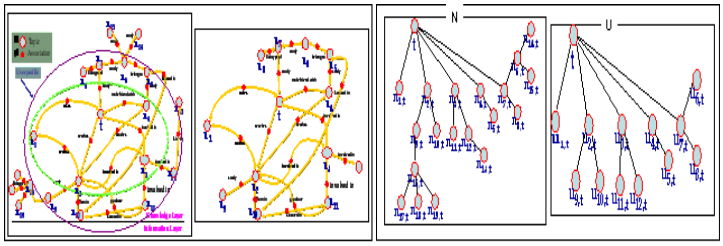


<표 1> 결과 및 순위

순 위	토 픽	가중치 값
1	토픽 "2"	0.063
2	토픽 "3"	0.047
3	토픽 "4"	0.044
	토픽 "7"	
5	토픽 "1"	0.041

본 논문에서는 관심 토픽에 관계된 토픽들의 가중치를 결정하기 위한 방법으로 그래프 구조를 트리 구조로 재편성하는 방법을 사용하였다.이는 PTR 알고리즘에서 관심 토픽을 기준으로 연결되는 토픽의 수가 어떤 의미를 가지는지 밝혀내기 위하여 관계의 성격을 제한한 것이다.

<그림 2> 관심 토픽을 중심으로 한 트리 구조로 재편성



$$W^2_{t,i} = Ti[\log\frac{(u_{i,t}+0.5)(N-n_{i,t}+0.5)}{(n_{i,t}+0.5)(U-u_{i,t}+0.5)}] \quad (2)$$

where, $n_{i,t}$: 토픽 맵 전체에서 관심 토픽 t에 대한 i번째 토픽과 관계되는 토픽 수

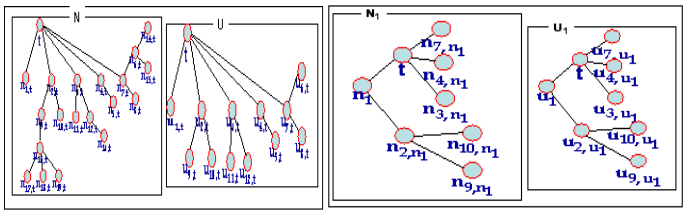
$u_{i,t}$: 사용자 프로파일 내에서 관심 토픽 t에 대한 i번째 토픽과 관계 되는 토픽 수

<그림 2>는 <그림 1>을 트리 구조로 재편한 것 인데우리는 수식 (1)의 방법과 동일하게 수식 (2)를 활용하여 가중치를 계산하였고, 결과 값은 $W_{t,1}$ = 0.038, $W_{t,2}$ = 0.044, $W_{t,3}$ = 0.044, $W_{t,4}$ = 0.041, $W_{t,7}$ = 0.044 로서 최종 순위는 토픽 2, 3, 7 - 4 - 1 순이다. 우리는 이 결과에서 토픽의 트리 구조 재편성이 관심 토픽의 관점에서 관계된 토픽들을 바르게 설명하고 있다고 할 수 있다.

2.2 관심 토픽과 관계된 토픽까지 확장

우리는 관계된 토픽의 중요 정도를 반영하기 위하여 관계된 토픽 별로 별도의 트리 구조를 형성하도록 하였다. 관심 토픽 t와 관련된 i번째 토픽을 중심으로 한 트리 구조의 예는<그림 3>과 같고, 이때 i번째 토픽에 연결된 j개의 토픽들의 수를 이용하여i번째 토픽의 중요 정도를 반영하는데 수식(3)과(4)를 이용하였다.

<그림 3> 토픽 "1" 을 중심으로 한 트리 구조화



$$n_{i,t'} = n_{i,t} + k\sum_j n_{j,ni}$$
$$u_{i,t'} = u_{i,t} + k\sum_j u_{j,ui} \quad (3)$$

$$W^3_{t,i} = Ti[\log\frac{(u_{i,t'}+0.5)(N-n_{i,t'}+0.5)}{(n_{i,t'}+0.5)(U-u_{i,t'}+0.5)}] \quad (4)$$

0< k ≤ 1의 범위를 가지는 k는 i 토픽의 중요도를 결정하는 상수이며 이때 k를 0.5(관계된 토픽의 2차 관계는 관심 토픽 t와의 관계보다 적게 산정되는 것이 타당하다)로 하여 가중치 $W^3_{t,i}$ 를 구한 결과, 토픽"2"와 토픽"7"은 0.060로 최상위 랭킹을 이루었고 토픽"3" 0.052, 토픽"4" 0.046, 토픽"1" 0.041의 순서로 랭킹을 구할 수 있었다.

3. 토 의

PTR 과 E-PTR 알고리즘을 이용하여 계산된 가중치와 순위를 비교해보면확장된 토픽이 많으면서도 사용자 프로파일과 원래의 토픽 맵과 그 형태가 유사한 토픽"7"이 상위 순위로 바뀐 것을 관찰할 수 있는데 이는 E-PTR 알고리즘이 PTR 알고리즘과 비교하여 사용자 프로파일의 해석에서 향상된 성능을 가지고 있음을 나타내는 것이다. 즉 원시 토픽 맵이 전문가의 노력에 의하여 형성된 바람직한 형태라면사용자가 특정한 토픽에 대한 선택 횟수가 증가하는 경우를 반영할 수 없는 토픽 맵의 가중 특성을 사용자의 관련 토픽에 대한 건전한 확장 정도를 이용하여 간접적으로 확인한 결과라고 할 수 있다. 또한 순위 간 가중치의 차이가 감소한 것을 알 수 있는데 이는 직접적인 관계의 수를 이용한PTR 알고리즘의 방법을 각 토픽의 중요도를 이용하여 완화한 결과로 보인다.

4. 결론 및 향후 연구

본 논문은 토픽 맵 환경 하에서 복합 관계를 이루고 있는 토픽들 간에 관계를 활용하여 사용자가 선호하는 주요 토픽들에 대한 관련 토픽의 순위 정보 제공을 목적으로 하였다.1차 관계를 이용하는 PTR 알고리즘에 비하여 E-PTR 알고리즘은 토픽 맵의 관계를 새로 해석하고 관계된 토픽들의 중요성을 반영함으로써 사용자의 의도에 근접한 토픽 순위를 제공할 수 있었다.

토픽 맵의 목적이 내재적인 요소를 해석하고자 하는 것인 만큼향후에는 알고리즘 적으로 보인 사용자 의도 근접에 대한 실제 사용자의 검증이 필요하며 이를 바탕으로 근접 정도의 반영 등 토픽 맵의 구현 방법의 변경과 관계(Association)와 같은 방법으로 어커런스를 이용한 토픽간의 관계를 규명하는 것 등을 계속적으로 연구하여야 하겠다.

참고 문헌

[1] 박정우, 김권일, 채진기, 이상훈, "사용자 프로파일을 이용한 개인화된 토픽 맵 랭킹 알고리즘 국방대학교 전산정보학과, 한국정보과학회 2007년 추계 학술 대회

[2] Steve Pepper, "The TAO of Topic Maps", Ontopia, 2006

[3] Karen Spärck Jones, Steve Walker, and Stephen E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (parts 1 and 2)". Information Processing and Management, pp.779-840, 2000.

[4] Stefan Buttcher, Charles L. A. Calrke, Brad Lushman, "Term proximity scoring for ad-hoc retrieval on very large text collections", Proceedings of the 29th annual international ACM SIGIR conference, Poster Session, pp.621-622, 2006.

[5] S. E. Robertson, S. Walker, "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval", Proceedings of the 17th annual international ACM SIGIR conference, pp.232-241, 1994.

[6] Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. "Okapi at TREC-7", In Proceedings of the Seventh Text Retrieval Conference. Gaithersburg, USA, November 1998.

[7] Jaime Teevan, Susan T. Dumais, Eric Horvitz, "Personalizing search via automated analysis of interests and activities" ,Proceedings of the 28th annual international ACM SIGIR conference, Session: User studies, pp.449-456, 2005.

[8] Feng Qiu, Junghoo Cho, "Automatic Identification of User Interest For Personalized Search", Proceedings of the 15th international conference on WWW, Session: Improved search ranking, pp.727-736, 2006.