

상대네트워크 구축에 의한 맞춤형 논문검색 시스템 모델링

User-oriented Paper Search System by Relative Network

Young Im Cho, Sang Gil Kang

경기도 화성시 수원대학교 IT대학 컴퓨터학과
E-mail: vchos@suwon.ac.kr, sgkang@suwon.ac.kr

요 약

이 논문은 사용자의 쿼리와 사용자의 행동양식을 바탕으로 상대네트워크를 구축함으로써 개인화된 논문검색 시스템을 모델링한 것이다. 제안하는 시스템은 사용자가 검색한 논문에서 키워드의 빈도수를 분석하여 개인적 상대네트워크를 구축하게 되는데, 이 네트워크는 다운로드, 열기, 삭제 등과 같은 사용자의 행동으로부터 키워드간 가중치를 조정을 함으로써 구축된다. 시스템의 성능평가를 위해 100명의 사용자들을 대상으로 실험한 결과, 기존의 검색엔진을 사용했을 때보다 성능이 우수하여 사용자 만족도가 높게 나타남을 알 수 있었다

Key Words : Relative network, User-oriented paper search system

1. 서 론

최근 정보의 홍수 속에서 연구자들은 자신들이 원하는 논문을 찾기가 쉽지 않다. 또한 사용자들이 검색엔진에 의해 찾은 논문에서 자신들이 원하는 정보를 찾기는 더욱 어렵다[1,2]. 만약 사용자들의 관심도를 반영한다면 논문검색시스템은 좀더 사용자의 쿼리에 가까운 논문을 검색할 수 있을 뿐 아니라 검색속도도 훨씬 빠를 것이다[3,4,5].

따라서 이 논문에서는 사용자의 쿼리 정보와 검색된 논문 리스트 상에서의 사용자의 행동과 선호도를 분석하여 사용자 맞춤형 상대네트워크를 구축하는 방법에 대하여 제안하고자 한다.

이렇게 구축된 상대네트워크에 의해서 검색된 논문에서의 사용자의 행동들, 예를 들면 '다운로드', '열기', '삭제' 등과 같은 행동들의 빈도수에 따라서 가중치를 부여하여 사용자 쿼리에 따른 맞춤형 상대네트워크를 구축함으로써, 사용자가 앞으로 논문 검색시 맞춤형으로 정보를 제공하고자 한다.

이 논문은 2장에서는 상대네트워크 모델링기법에 대한 주요 알고리즘을 소개하고 3장에서는 상대네트워크 모델링에 의한 논문검색시스템의 전체 개요를 설명하고 4장에서 실험결과를 통해 5장에서 결론을 제시하고자 한다.

2. 상대네트워크 모델링 기법

2.1 상대네트워크 모델링

키워드 기반 사용자의 쿼리로 부터 논문들이 검색되면, 이들 검색된 논문의 요약문에서 요약문에 있는 명사와 키워드간의 상대적 관계를 측정할 수 있다는 것이 상대네트워크의 주요 포인트이며 본 논문에서 제안하는 중점사항이 된다.

만약 사용자가 검색엔진에 키워드(명사)를 갖고 쿼리를 하면, 검색엔진은 동일 키워드를 갖는 관련된 논문을 모두 검색한다. 이로부터, 파싱에 의해 다음 식(1)에서와 같이 이 키워드(명사)가 논문에 몇 번 나왔는지를 측정한다.

$$freq(keyword, noun_i) = \sum_{a=1} n_{i,a} \quad (1)$$

여기서 $n_{i,a}$ 는 a 번째 검색된 문서에 출현하는 명사 i 의 반복횟수를 말한다.

$freq(keyword, noun_i)$ 는 전체 요약문에서 키워드와 관련되어 검색된 명사 i 의 상대적 출현 정도를 나타내는 강도(strength)를 의미한다. 키워드와 각 명사에 대한 상대적 강도는 다음 그림 1에서와 같이 네트워크 형태(본 논문에서

는 키워드 상대네트워크 또는 상대네트워크라 명명)로 모델링 됨을 본 논문에서 제안한다.

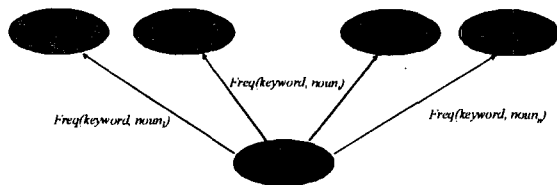


그림 1. 키워드 상대 네트워크

이 그림에서 보는 바와 같이 키워드와 명사간의 상대적 강도는 명사가 요약문에 출현하는 빈도수에 의해 측정된다. 만약 명사의 출현횟수가 많으면 이 명사는 키워드와 관련성이 매우 높은 것으로 간주한다. 같은 방법으로 모든 명사는 사용자에게 의해 쿼리된 가능한 키워드일 수 있다. 즉, 키워드와 논문에 출현하는 명사와의 관계로부터 상대네트워크를 구축하여 사용자 맞춤형으로 정보를 검색해 주는 것이다.

이 개념을 좀더 일반화 시켜보면 키워드와 명사간의 전적인 다중 네트워크(fully connected network)형태의 구축이 가능하다.

2.2 맞춤형 상대네트워크 모델링

상대네트워크에 의해 구축된 것으로부터 개인 맞춤형 논문검색을 위해서는 다음 식(2)에서와 같이 사용자의 행동들을 바탕으로 검색한 후, 이것들을 필터링하여 사용자에게 결과를 인터페이스를 통해 제시할 수 있다. 맞춤형에 관해서는 [6,7,8]논문에서 제시되어 있다.

$$freq(keyword_i, keyword_j) = \sum_{a=1} w_a n_{i,j,a} \quad (2)$$

이 식에서 w_a 는 논문 a 에 대한 사용자의 행동의 가중치를 말하며, $n_{i,j,a}$ 는 쿼리 i 에 의해 a 번째 검색된 논문으로부터 키워드 j 의 출현 빈도수를 의미한다. 일반적으로 '다운로드'된 논문은 사용자가 매우 관심있는 논문으로 간주하고 '열기'를 한 논문은 관심도가 비교적 높고, 아무런 행동을 하지 않은 논문은 사용자가 관심이 없는 것으로 간주하여 이들 논문에서 출현하는 명사와 사용자 쿼리에 해당하는 키워드와의 상대적 관련성을 네트워크로 구축한다.

그러나 하나 이상의 키워드들과 사용자 쿼리와의 관련성을 나타내기 위해서는 식(2)를 일반화 시켜야 하는데 이것을 나타낸 것이 식(3)이다.

$$freq(Set_keyword, keyword_j) = \sum_{k=1} freq(keyword_k, keyword_j) \quad (3)$$

이 식에서는 키워드 집합에 대하여 키워드 j 의 상대적 강도를 나타내고 있다. 여기서 $Set_keyword$ 는 쿼리된 키워드의 집합을 의미하며, s 는 쿼리된 키워드의 수를 의미한다. 식(3)은 좀더 식(4)와 같이 정규화된 형태로 나타낼 수 있다.

$$Nfreq(Set_keyword, keyword_j) = \frac{freq(Set_keyword, keyword_j)}{\sum_{j=1} freq(Set_keyword, keyword_j)} \quad (4)$$

여기서 $Nfreq(Set_keyword, keyword_j)$ 는 키워드 집합에 대해서 키워드 j 의 정규화된 상대적 강도를 의미한다. 정규화된 상대적 키워드의 값에 따라서 리스트에 있는 검색된 논문들의 순서가 결정된다. 예를 들면, 쿼리된 키워드를 사용하는 검색된 논문들로부터 높은 상대적 강도를 갖는 논문들이 높은 값으로 랭킹되어 있다. 즉, 검색된 페이지의 앞부분에 나타나게 된다.

3. 사용자 맞춤형 논문검색시스템

3.1 전체 시스템 구조

본 논문에서 제안하는 상대네트워크를 이용한 사용자 맞춤형 논문검색시스템의 전체구조는 다음 그림 2와 같다.

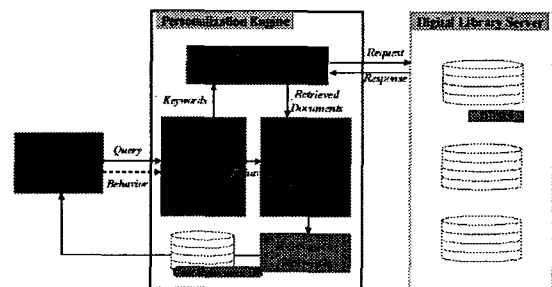


그림 2. 전체 시스템 구조

이 시스템은 논문수집모듈(GM), 검색모듈(RM), 개인화모듈(PM)에서와 같이 3가지 모듈로 구성되어 있다. GM은 쿼리된 키워드와 사용자 인터페이스를 통한 행동들을 수집한다. 쿼리된 키워드는 RM을 통해 디지털 라이브러리 서버로 이동한다. 디지털 라이브러리 서버는 DBMS로부터 키워드를 이용하여 관련된 논문들을 검색하여 RM으로 보낸다, RM은 논문들로부터 요약문을 검색하여 파싱에 의해 키워드를 PM에게 보낸다. PM은 사용자의 행동들에 따라 키워드는 사용자의 개인화된 상대네트워크를 모델링하기 위해 사용된다. 또한 사용자의 행동들과 그들의 상대네트워크들은

사용자 관련성 DB(user relevance DB)에 저장된다. 사용자 관련성 DB는 사용자 ID, 키워드 리스트, 키워드간의 상대적 강도 등으로 구성되어 있다. 이러한 개인 맞춤형 상대네트워크에 의해서 사용자에게 다른 논문검색결과는 사용자 인터페이스를 통해서 보여지게 된다.

4. 시뮬레이션

이 시스템의 성능평가를 위해서 윈도우 NT 기반에서 Visual C++을 이용하여 사용자 관련성 DB를 구축하여 시스템의 성능평가하였다. 실험을 위해 식 (2)에서와 같이 논문 *a*에 대한 사용자의 행동에 대한 가중치 값을 선택한다. 먼저 다운로드에 대한 가중치는 $w_a = 5$ 이고, 열기에 대한 가중치는 $w_a = 2$ 이고, 아무런 행동을 하지 않은 것에 대한 가중치는 $w_a = 1$ 로 한다. 또한 키워드여부에 대한 출현빈도수의 쓰레숄드를 2로 한다. 이 값들은 수원대학교에서 2005년 1월부터 2006년 3월까지 100명의 사용자들을 대상으로 실험을 해서 정량적으로 얻은 최적화된 값들이다.

기존 검색엔진에 의한 검색결과와 본 논문에서 제안하는 사용자 상대 네트워크에 의한 키워드검색에 의한 검색결과를 비교하기 위해 ACM Portal 사이트의 검색결과와 제안하는 시스템의 검색결과를 비교해 보고자 한다.

검색결과, 'network'라는 키워드를 ACM 포털에서 검색한 결과 64,493개의 논문들이 검색되었다. 이것들로부터 사용자가 쉽게 관심있는 논문들을 참고하기란 쉽지 않다. ACM 포털에서 사용한 것과 동일한 키워드 'network'를 입력한 결과, 그림 3에서와 같이 'network'와 관련성이 가장 높은 0.47의 상대적 강도를 갖는 'neural'이란 키워드를 갖는 논문들이 다수 검색되었다. 즉, 'network'와 'neural'이 동시에 들어간 논문들이 다수 검색되어 상위 랭크되어 있었다.

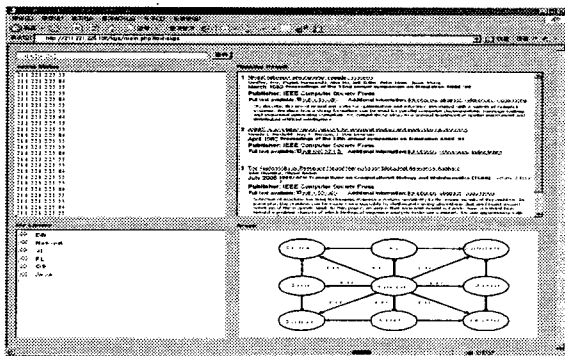


그림 3. 제안된 시스템의 논문검색결과

즉, 사용자가 선호하는 단어들간의 연관성을 고려하여 상대네트워크에 의해 논문들이 검색되었다.

5. 결과

이 논문에서는 사용자의 쿼리 키워드와 검색된 논문들로부터 사용자의 행동을 바탕으로 키워드와 검색된 논문들간의 명사들간의 상대네트워크를 구축하여 사용자 관련성 DB를 구축함으로써, 향후 맞춤형 정보검색이 가능하도록 제안한 논문이다.

이 논문에서 제안한 시스템을 구현하여 사용자 관련성 DB를 구축한 결과, 사용자의 쿼리된 키워드와 상대적 관련성이 높은 명사들을 갖고있는 논문들이 상위 랭크되어 제시되어 사용자의 만족도가 향상되었다.

참 고 문 헌

- [1]5. Ashrafi, M.Z., Tanizr, D., Smith, K., "ODAM: An Optimized Distributed Association Rule Mining algorithm," IEEE Distributed Systems Online, vol. 3, no. 3, pp. 1-18, 2004
- [2] Ciaramita, M., Johnson, M., "Explaining away ambiguity: Learning verb selectional preference with Bayesian networks," Proc. Intl. Conference on Computational Linguistics, pp. 187-193, 2000
- [3] Maltz, D.A., "Distributing Information for Collaborative Filtering on Usenet net News," SM Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1994
- [4] Linden, G., Smith, B., York, J., "Amazon.com Recommendations: Item-To-Item Collaborative Filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, 2003
- [5] Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T., "Evaluating Collaborative Filtering Recommender Systems", ACM Transactions on Information Systems vol. 22, no. 1, pp. 5-53, 2004
- [6] Cotter, P., Smyth, B., "Personalization Techniques for the Digital TV world," Proc. European Conference on Artificial Intelligence, pp. 701-705, 2000
- [7] Lee, W.P., Yang, T.H., "Personalizing Information Appliances: A Multi-agent Framework for TV Program Recommendations," Expert Systems with Applications, vol. 25, no. 3, pp. 331-341, 2003
- [8] Torres, R., McNee, S.M., Abel, M., Konstan, J.A., Riedl, J., "Enhancing Digital Libraries with TechLens+", ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 228-236, 2004