블로그 추천을 위한 내용 유사 클러스터 기반의 블로그 평가

김현정*, 김무철*, 한상용*
*중앙대학교 컴퓨터공학과
e-mail : hyunjung.kim@ec.cse.cau.ac.kr

Evaluation Method based on Contents and Social Network for Blog Recommendation

Hyun-Jung Kim*, Mu-Cheol Kim*, Sang-Yong Han*
*Dept. of Computer Engineering, Chung-Ang University

요 약

본 연구는 최근 블로그 추천 연구의 주요 쟁점으로 제기되는 추천 후보의 선정과 추천 후보 평가에 접근한다. 첫 번째로 추천 후보 선정은 추천 요구자와 소셜 네트워크 관계에 있는 블로그를 중심으로 진행한다. 이러한 접근방식은 추천 요구자가 타 블로그와 직접적인 관계를 많이 이루지 못했을 경우 다수의 간접 연결 블로그가 추천 후보로 차지하게 된다. 직접 관계의 희소함으로 인하여 추천 후보와 추천 요구자와의 연관성이 전체적으로 저하되는 문제에 초점을 맞추어 추천 대상을 내용 기반의 클러스터 단위로 선정하는 방식을 제안한다. 또한 추천 대상 블로그의 평가에서는 소셜 네트워크 및 내용 평가를 결합시킴으로써 요구자에게 보다 적합한 추천 결과를 제시한다.

1. 서론

기하급수적인 블로그의 성장은 블로거 간의 의미 있는 관계형성과 블로그 정보수집을 어렵게 만든다. 이러한 현상 가운데 블로그 추천 연구는 각 개인에게 보다 적합한 정보를 제공함으로써 블로그를 효율적으로 관리하려는 목적으로 진행되었다.

최근 블로그 추천 연구의 주요 쟁점은 추천 후보 선정과 추천 후보 블로그의 평가이다. 첫 번째로 추천 후보 선정은 추천 요구자와 소셜 네트워크 관계에 있는 블로그를 이용한다. 이 과정에서 추천 요구자가 타 블로그와 직접적으로 많은 관계를 이루지 못했을 경우 추천 후보에 다수의 간접 연결 블로그가 차지한다. 직접 관계의 희소함으로인해 추천 후보와 추천 요구자와의 연관성이 전체적으로 저하되는 희소성 문제[1]는 최종적인 블로그 추천 결과를 약화시킬 수 있기 때문에 문제해결의 중요성이 강조되어야 한다. 다음으로 제기되는 쟁점은 추천 후보를 대상으로 추천 결과를 도출하기 위한 블로그의 평가이다. 개인에게 맞는 추천을 위한 블로그 평가 요소로써 많은 연구들이 블로그 간 연결도 및 내용 유사도를 이용했다. 각 평가 요소들은 개별적으로 혹은 결합적으로 적용될 수 있는데 블로그 평가 결과의 적합성을 보다 높이기 위해서 다양한 작업들이 요구된다.

본 연구에서는 추천 후보 선별과정에서 추천 요구자와 단순 연결로 생성된 기본 블로그 데이터들의 내용 유사성을 이용해 클러스터링을 한다. 클러스터 단위로 추천 후보군을 선별함으로써 블로그 추천에서 나타나는 희소성 문제에 접근한다. 이후 선정된 추천 후보군을 통해 블로그 평가의 정규화 과정을 진행하고 최종적으로 블로그 평가 결과의 적합성을 높인다.

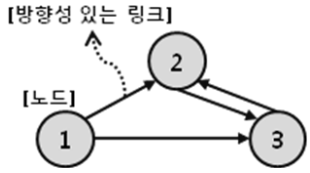
본 논문의 구성은 다음과 같이 모두 5 장으로 구성되어 있다. 2 장에서는 블로그 추천과 관련된 연구들을 분석한다. 또한 3 장에서는 내용 유사 클러스터링을 통한 블로그 추천 대상의 선정 및 확장 방식을 제안하며, 4 장에서는 결정된 추천 대상을 바탕으로 블로그를 평가한다. 마지막으로 5 장에서는 결론과 함께 앞으로의 연구 방향을 기술한다.

2. 관련 연구 및 연구 배경

블로그 사용자에게 유용한 정보를 제공하기 위해서는 블로그가 가진 속성으로 평가기준을 정해서 각 블로그들을 평가해야 한다. 블로그 평가를 위한 주요 속성으로는 블로그간의 연결 관계와 블로그 문서 내용이 있다[5][6].

2.1. 소셜 네트워크에 기반한 평가 모델

블로그를 이용하는 개인은 타 블로거가 제공하는 정보의 수용자이자 또 다른 정보의 제공자가 될 수 있다. 블로거가 가지는 이중적 역할구조를 통해 블로거는 서로간에 연결고리를 만들고 확장하면서 다양한 관계 구조를 생성해낸다[2]. 이러한 관계구조는 (그림 1)과 같이 한 명의 블로거를 노드로, 블로거 간 연결 관계를 링크로 표현하는 소셜 네트워크에 접목시킬 수 있다. 소셜 네트워크에서의 링크는 방향성을 가질 수 있으며 부여하는 의미에 따라서 다양한 적용이 가능하다[3].



(그림 1) 노드와 링크로 구성된 소셜 네트워크

소셜 네트워크를 이용한 블로그 평가는 일반적으로 링크가 가지는 방향 및 개수 속성을 통해 이루어진다. [4]에서는 블로그 간의 연결을 인링크(In-link)와 아웃링크(Out-link)로 구분해서 각 블로그의 역할 및 영향력을 판단하는 모델을 제시했다(그림 2).

$$Authority(v) = \sum_{v \in S, v \to p} Hub(p) \quad \text{AND} \quad Hub(p) = \sum_{v \in S, p \to v} Authority(v)$$

(그림 2) 블로그 영향력 모델

2.2. 내용기반의 평가 모델

블로그 사용자는 자신의 관심 주제를 포함하고 있는 블로그에 접근하려는 성향을 지니고 있다. 이러한 성향은 블로그 문서의 내용이 유사한 블로그끼리 연결될 가능성이 높음을 의미한다[6]. 따라서 블로그 사용자들의 문서내용이 어느 정도 유사한지를 비교, 평가함으로써 사용자가 원하는 정보를 제공할 수 있다.

내용 기반의 평가를 위해서는 블로그 문서에 포함된 여러 단어들을 이용해 비교, 분석하게 된다. [7]에서는 대표 문서 페이지(Seed page)로부터 키워드를 추출한 다음 새로운 페이지가 키워드를 어느 정도 포함하고 있는지를 판단함으로써 대표 문서 페이지와 가까운 블로그를 선별한다. 평가를 위한 단어로는 정보의 주제뿐만 아니라 의견을 나타내는 단어도 이용할 수 있다[8].

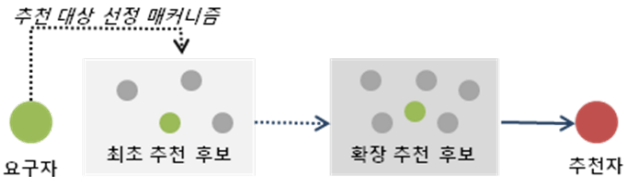
2.3. 통합 평가 모델

내용기반과 소셜 네트워크 기반의 상이한 블로그 평가 방식은 반드시 독립적으로 시행되어야 하는 것은 아니다. 오히려 두 방식을 통합함으로써 블로그 평가 결과를 더욱 향상될 수 있다[9][10].

3. 블로그 추천대상 선정 및 확장

3.1. 블로그 추천의 설정 범위

본 연구의 블로그 추천 환경에서는 블로그 추천의 주체를 (그림 3)과 같이 블로그 추천을 받고자 하는 요구자와 그의 추천 대상이 되는 추천자로 정의한다.



(그림 3) 블로그 추천 과정

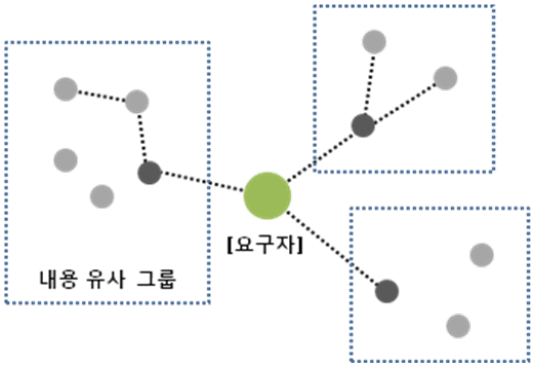
추천 요구자는 한 명의 블로거로 나타내며 요구자의 블로그에 포함된 프로파일을 바탕으로 추천 후보를 선정하고 확장시킨다. 확장된 추천 대상은 블로그 추천 결과를 위한 최종적인 비교 대상이 된다.

3.2. 최초 추천후보의 선정을 위한 내용 유사 기반 클러스터링

최초 추천후보는 요구자와 연결된 블로그를 중심으로 선정된다. 블로그들의 연결관계는 소셜 네트워크에서 FOAF(Friend of a Friend)[3]와 같이 직접관계와 간접관계로 이루어져 있는데 직접관계는 대부분 소수의 블로그에 몰려있다. 이는 다른 의미로 많은 블로그들이 최초 추천후보를 간접관계의 블로그에서 선정함을 뜻한다. 직접 관계가 희소한 추천후보는 추천 요구자와의 연관성이 전체적으로 저하될 수 있으며 이는 희소성 문제(Sparsity Problem)[1]로 정의된다. 희소성 문제는 블로그 추천의 결과를 약화시키는 주요 원인이 될 수 있다.

본 연구에서는 희소성 문제를 해결하기 위해 직접 및 간접 연결 블로그를 내용 유사성을 통해 집단으로 구성하고 각 집단과 추천 요구자를 비교함으로써 추천 후보를 선정한다.

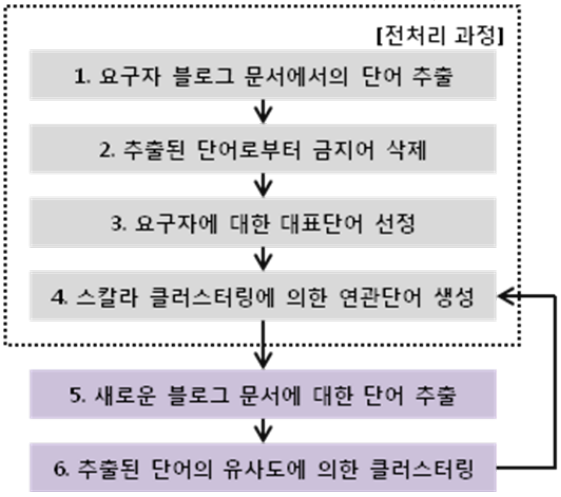
이러한 희소성 문제 해결과정을 구체적으로 살펴보면 크게 두 단계로 나누어 진행된다. 첫 번째 단계에서는 요구자의 블로그로부터 직, 간접 연결로 생성된 다수의 블로그들을 서로 간에 관계를 따지지 않고 내용 유사도만으로 클러스터링한다(그림 4). 이는 희소성 문제에 대한 최소한의 작업을 의미하는데, 요구자와의 연결성이 낮은 블로그인 경우에도 내용적인 측면을 고려해 최초 추천 후보로 참여할 수 있는 기회를 제공하고자 한다.



(그림 4) 요구자 중심의 클러스터링 구조

내용 유사 기반의 클러스터링을 위해서는 (그림 5)의 점선 박스로 표시된 부분과 같이 요구자의 블로그에 대한 전처리 과정을 가진다. 전처리 과정은 요구자와 가까운 대표단어들을 미리 선택하고 분류, 확장함으로써 새로운 블로그에 대한 효율적인 클러스터링을 진행하려는 목적을 가진다. 이러한 전처리 과정을 위해서 스칼라 클러스터링에서 제시하는 방식을 적용한다. 스칼라 클러스터링[11]은 키워드 간의 연관성을 측정하기 위해 각 키워드가 비슷한 이웃을 가지면 서로 연관되어 있다고 판단하게 된다. 전처리 과정에서 스칼라 클러스터링 기법을 적용하게 되면 요구자의 블로그 문서로부터 대표단어와 이웃단어들이 각각 기본 클러스터를 생성한다.

요구자 블로그 기준의 전처리 과정이 종료되면 이후 직, 간접적으로 연결된 새로운 블로그들을 기본 클러스터에 기준해서 또 다시 클러스터링 과정을 진행한다. 각각의 기본 클러스터와 새로운 블로그 문서를 개별적인 벡터값으로 정하고 벡터들의 코사인 유사도[11]를 통해 추천 후보 선정을 위한 최종적인 클러스터가 완성된다. 각 클러스터들은 최종적인 블로그 평가를 위해서 반드시 직접연결 블로그를 하나 이상 포함해야 하기 때문에(그림 4) 직접연결 블로그를 가지지 못한 클러스터가 발생했을 경우 내용적으로 가장 유사성을 보이는 클러스터와 결합하게 된다.



(그림 5) 내용 유사도에 의한 클러스터링 과정

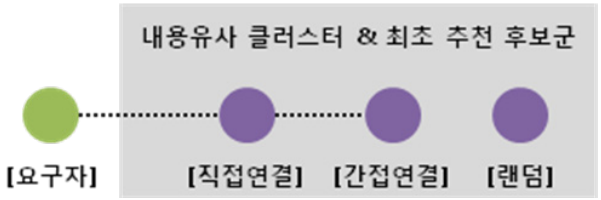
3.3. 최초 추천 후보군 선정 및 확장

최초의 추천 후보는 앞 절에서 설명한 내용 유사 기반의 클러스터 단위로 선정되는데 이는 희소성 문제와 관련이 있다. 희소성 문제는 간접 연결의 블로그가 일반적으로 추천 요구자와의 연관성이 낮기 때문에 발생한다. 그렇기 때문에 간접 연결 블로그가 직접 연결 블로그와 내용 유사성으로 묶여질 수 있었다면 추천 후보 선정을 위한 요구자와의 연관성 판단 또한 서로 묶인 블로그끼리 함께 진행되어야 한다.

최종적으로 최초 추천 후보군을 선출하기 위해 각 클러스터들은 요구자와의 평균 소셜 네트워크 연결도 및 평균 내용 유사도를 판단하게 된다. 내용 유사 클러스터 단위에 의해 정규화된 값들은 추천 요구자와의 연결 확률 수치를 보다 신뢰하게 한다. 클러스터 단위로 각 수치들의 비교가 끝나면 비교결과가 상위에 속하는 클러스터들은 최초 추천대상으로 선정된다.

최초 추천후보군 선정이 끝나게 되면 사용자 정의의 랜덤 가중치를 통해 추천 후보 대상이 보다 확장될 수 있다. 최초 추천 후보군에 포함된 3 종류의 블로그(그림 6) 중 어떤 블로그와도 연결되지 못한 랜덤 블로그가 해당 가중치에 영향을 받는다. 랜덤 블로그가 원래 추천 요구자와 간접적으로 연결된 블로그이다. 하지만 내용 유사성에 의해 클러스터가 결정됨으로써 이전 연결이 끊어지게 되어 랜덤 블로그가 발생하게 된 것이다. 비록 클러스터 내에서 타 블로그와 연결관계를 가지지 않더라도 요구자와의 연관성을 만족시켰기 때문에 랜덤 블로그를 통해 새로운 연결관계를 생성하는 것은 추천 후보 확장에도 충분한 가치를 가진다.

이로써 요구자와의 연관성을 지속적으로 고려한 추천 후보의 선정 및 확장과정 통해서 다수의 간접 연결이 추천 후보에 속하는 상황으로 인해 발생하는 희소성 문제를 근본적으로 해결할 수 있게 된다.



(그림 6) 최초 추천 후보군의 구성

4. 블로그 평가

4.1. 블로그 평가기준

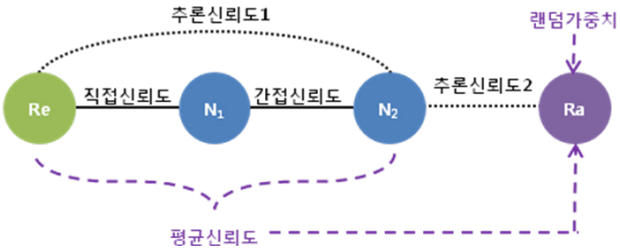
추천을 위한 후보 블로그가 결정이 되면 평가 요소에 따라서 각각의 블로그를 평가하게 된다. 평가 요소는 크게 소셜 네트워크 평가와 내용 평가로 나뉘게 된다.

소셜 네트워크 평가는 평가범위에 따라 다시 신뢰도와 평판 평가로 세분화된다. 신뢰도(Trust) 평가는 요구자와의 소셜 네트워크 연결성을, 평판(Reputation) 평가는 전체 참여한 추천 후보를 대상으로 소셜 네트워크 연결성을 판단하게 된다.

내용 평가는 요구자가 제시하는 블로그 정보와 내용적인 유사성을 판단함으로써 요구자가 보다 원하는 정보를 찾기 위한 요소로 작용된다.

4.2. 신뢰도 평가

신뢰도는 요구자와의 주관적인 관계를 평가하려는데 그 목적이 있다. 추천 후보군을 이루는 구성원은 직접연결, 간접연결, 랜덤연결로 이루어져 있다(그림 4). 직접 연결 블로그에서는 요구자의 블로그에서 실제적으로 이루어진 소셜 네트워크 활동을 측정함으로써 신뢰도를 평가하게 된다. 또한 간접연결 블로그는 직접연결 블로그의 신뢰도를 바탕으로 추론된 결과가 신뢰도로 적용된다. 마지막으로 랜덤연결 블로그의 신뢰도는 해당 클러스터의 직접 및 간접 연결 신뢰도의 평균값을 기준으로 앞서 블로그 추천후보 확장 때 정의된 랜덤 가중치에 따라 신뢰도가 결정된다(그림 7).



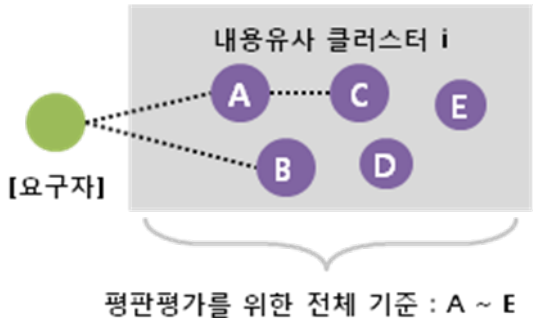
(그림 7) 신뢰도 추론

4.3. 평판 평가

신뢰도가 요구자와 개인적인 관계를 평가하는 요소

라면, 평판은 전체를 대상으로 구성원의 객관적인 평가를 위한 척도이다.

평판 평가에서 가장 중요한 사항은 각 추천 후보들의 객관적 지표가 될 전체의 기준을 설정하는 것이다. 만약 크롤링된 모든 블로그를 대상으로 평판 평가가 이루어 진다면 연관성이 없는 정보 범위를 가진 블로그들이 모여 평가가 진행된다. 이는 결과적으로 블로그 사용자들이 상대적으로 관심을 가지지 않는 주제를 제공하는 블로그는 불리한 결과를 받을 수 있다. 이러한 문제를 해결하기 위해 전체 대상의 기준을 (그림 8)와 같이 각 블로그가 속해있는 내용 유사 클러스터 로 정한다. 내용 유사 클러스터는 클러스터링 과정에서 내용적으로 관련을 가지는 블로그끼리 엮어 있기 때문에 평판 평가의 정규화를 위한 이상적인 설정이라고 볼 수 있다.



(그림 8) 평판평가를 위한 전체 기준의 설정

4.4 내용 평가

앞서 신뢰도 및 평판 평가가 블로그에서의 소셜 네트워크를 이용했다면 마지막 평가 요소인 내용 평가는 추천 후보와 요구자의 블로그 내용을 직접적으로 비교, 평가한다.

우리는 내용 평가 이전에 추천 후보군을 선정하기 위해서 내용 유사 클러스터링을 통해 이미 내용적 측면으로 블로그를 판단하고 분류했다. 하지만 클러스터링 과정에서는 요구자와 각각의 블로그가 개별적으로 내용 유사도는 비교하지 않았기 때문에 최종 추천 결과를 위해 내용 평가를 진행하게 된다.

5. 결론 및 향후 연구

본 연구에서는 최근 블로그 추천 연구의 주요 쟁점으로 제기되는 추천 후보의 선정과 추천 후보 평가에 접근하였다. 소셜 네트워크 관계로 생성된 기본 블로그 데이터에는 다수의 추천 요구자들의 직접 연결이 적음으로 인해 대부분 다수의 간접연결로 구성되었다. 이러한 기본 블로그 데이터의 성향때문에 추천 후보와 추천 요구자와의 내용 연관성이 전체적으로 저하되는 희소성 문제가 발생하여 기본 블로그 데이터를 내용 유사도에 클러스터링하여 클러스터 단위로 추천 후보를 선정했다. 또한 추천 대상 블로그의 평가에서는 소셜 네트워크 및 내용 평가를 결합시킴으로써 요구자에게 보다 적합한 추천 결과를 제시할 수 있었다.

향후에는 블로그 평가에서 평판뿐 아니라 신뢰도 및 내용 평가 또한 정규화 과정을 거침으로써 블로그 평가 결과값을 보다 정확하게 구현하고자 한다.

Acknowledge

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the HNRC(Home Network Research Center) - ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)(NIPA-2009-C1090-0902-0035) and also partially supported by National Research Foundation of Korea Grant funded by the Korean Government(2009-0076290).

참고문헌

[1] Zan Huang, Daniel Zeng, Hsinchun Chen, “A Link Anaysis Approach to Recommendation under Sparse Data”, AMCIS, 2004.

[2] 레베카 블러드, “블로그 : 1 인 미디어시대”, 전자신문사, 2003.

[3] Sun-hong Kim, Sang-yong Han, "The Method of Inferring Trust in Web-based Social Network using Fuzzy Logic", MIR Labs, 2009.

[4] Akshay Java, Pranan Kolari, Tim Finin, Tim Oates, “Modeling the Spread of Influence on the Blogosphere”, WWW, 2006.

[5] 노기영, 이미영, “블로그의 매체경쟁에 관한 연구 : 관계지향 블로그와 정보지향 블로그의 적소분석을 통한 경쟁분석”, 2005.

[6] Kyumars Sheykh Esmaili, Mahmood Neshati , Mohsen Jamali, Hassan Abolhassani, Jafar Habibi, “Comparing Performance of Recommendation Techniques in the Blogsphere”, ECAI, 2006.

[7] Chumki Basu, Haym Hirsh, and William W. Cohen, “Recommendation as classification : Using social and content-based information in recommendation”, AAAI/IAAI, 1998.

[8] BenHe, Craig Macdonald, Iadh Ounis, “Ranking Opinionated Blog Posts using OpinionFinder”, SIGIR, 2008.

[9] Yung-Ming Li, Ching-Wen Chen, “A Synthetical Approach for Blog Recommendation : Combining Trust, Social Relation and Semantic Analysis”, Elsevier, 2008.

[10] Tse-Ming Tsai, Chia-Chun Shih, Seng-cho Tl. Chou, “Personalized Blog Recommendation Using the Value, Semantic and Social Model”, Innovations in Information Technology, 2006.

[11] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, “Modern Information Retrieval”, Addison Wesley, 1999.