PARAFAC 분해를 이용한 블로그 공간 분석

김기남, 김상욱, 김진우
한양대학교 전자컴퓨터통신학과
e-mail: {kinam, wook, racerkim86}@agape.hanyang.ac.kr

An Analysis of a Blogosphere using PARAFAC Decomposition

Ki-Nam Kim, Sang-Wook Kim, Jin-Woo Kim
Department of Electronics and Computer Engineering, Hanyang University

요 약

본 논문에서는 블로그 공간을 텐서로 표현하고, 이를 분석한다. 분석 결과에 따르면, PARAFAC 분해를 통하여 특정 주제를 나타내는 커뮤니티들을 올바르게 파악할 수 있었으며, 각 커뮤니티에서 영향력 있는 블로그들과 키워드들, 그리고 권위 있는 포스트들을 식별할 수 있었다.

1. 서론

블로그 공간은 대표적인 온라인 사회연결망 중 하나이다. 블로그는 블로그의 주인인 블로거가 자신의 의견이나 생각을 글로써 온라인상에 저장할 수 있는 일종의 개인 홈페이지이다[1]. 블로거는 임의의 포스트 q와 연관된 내용의 포스트 p를 작성하고, p에 q에 대한 링크를 부여하는 행동인 트랙백(trackback)과 다른 블로그에 존재하는 포스트를 자신의 블로그로 복사해오는 행동인 스크랩(scrap)을 할 수 있다[2]. 이 두 가지 행동은 특정 포스트에 대한 블로거의 관심을 나타낸다. 따라서 이러한 행동들은 블로그 공간에서 일어나는 다양한 활동을 분석하는데 이용될 수 있다.

블로그 공간과 관련된 가장 흥미 있는 문제로는 (1) 커뮤니티 식별, (2) 커뮤니티 안에서 권위 있는 포스트식별, (3) 커뮤니티 안에서 영향력 있는 블로그 식별 등이 있다. 본 논문에서는 블로그 공간을 3차원 텐서(tensor)로 표현하고 대표적인 텐서 분해 기법인 PARAFAC 분해[3]를 이용하여 세 가지 문제를 동시에 해결한다. 또한, 해당 기법이 커뮤니티 식별과 각 커뮤니티 내에서 영향력 있는 블로그와 키워드, 그리고 권위 있는 포스트를 찾아내는데 유용하다는 것을 규명한다.

2. PARAFAC 분해를 이용한 분석

본 논문에서는 실험을 위해 국내 블로그 사이트로부터 수개월간 수집하여 익명으로 처리한 데이터를 사용한다. 각 포스트의 제목으로부터 추출한 다수의 형태소들을 해당 포스트의 키워드로 사용하며, 더 정확한 실험을 위해 'a', 'the' 같은 노이즈 단어들을 제거하였다. PARAFAC 분해를 이용하여 블로그 공간을 분석하기 위해서 블로그 공간을 $N \times M \times K$ 3차원 텐서 B로 변환한다. 이때, N은 블로그의 수, M은 포스트의 수, K는 키워드의 수이다. 블로그 i와 포스트 j사이에 스크랩이나 트랙백이 존재하고 해당 포스트의 제목이 키워드 k를 포함하는 경우, 텐서 B의 요소인 B_{ijk}는 1의 값을 갖는다. 그렇지 않은 경우, B_{ijk}는 0의 값을 갖는다. PARAFAC 분해를 이용하면 텐서B는 식 1과 같이 표현된다.

$$B \approx \sum_{r=1}^{R} (\lambda_r u_r \circ v_r \circ w_r) \qquad \text{(식 1)}$$

벡터 u_r, v_r, w_r는 r 번째 커뮤니티에 대한 블로그, 포스트, 그리고 키워드의 관련 정도를 각각 나타낸다. 따라서 커뮤니티 안에서 높은 점수를 받은 블로그, 포스트, 그리고 키워드들은 해당 커뮤니티에서 상호 관계가 있는 것으로 해석된다.

표 1은 PARAFAC 분해의 변수인 R이 변화함(10, 30, 50)에 따라 도출되는 커뮤니티들의 주제를 나타낸다. R은 사용자에 의한 입력 값으로 본 실험에서는 찾고자하는 커뮤니티의 수를 의미한다. 표에서 각 열은 실험에 사용된 R과 도출된 커뮤니티들의 주제들이다. 주제(i)에서 i는 해당 주제에 속하는 커뮤니티 수를 의미한다. R이 증가함에 따라서 도출되는 커뮤니티의 수가 증가하였으며, 새로운 주제의 커뮤니티들이 도출되었다.

좀 더 자세한 분석을 위해서, 각 커뮤니티 내에서 높은 점수를 받은 상위 4개의 키워드, 영향력이 높은 상위 4개의 블로그에서 출현 빈도가 높은 4개의 키워드, 그리고 권위 높은 포스트 상위 4개의 제목이 가지고 있는 키워드를 비교하였다.

표 2는 도출된 커뮤니티들 중에서 첫 번째 커뮤니티에 대한 분석 결과를 나타낸다. 동일 커뮤니티에 있는 높은 점수를 받은 상위 키워드들과 포스트들의 제목이 가지고

있는 키워드들 그리고 출현 빈도가 높은 키워드들이 일치하는 것을 확인할 수 있다.

<표 1> PARAFAC 분해를 이용하여 도출된 커뮤니티들의 키워드.

R	도출된 커뮤니티
10	리폼(2), 종이(2), 일본어(1), 영어(2), 액세서리(1), 습관(1), 기능(1)
30	리폼(3), 종이(5), 일본어(2), 영어(9), 액세서리(2), 습관(1), 기능(1), 혈액형(1), 성격(1), 요리(4), 속담(1)
50	리폼(7), 종이(7), 일본어(2), 영어(13), 액세서리(1), 습관(4), 기능(1), 혈액형(1), 성격(2), 요리(5), 속담(1), 아이디어(1), 영화(1), 다이어리(1), 별자리(1), 광고(1), 토익(1)

본 논문에서는 PARAFAC 분해를 이용하여 도출된 커뮤니티들의 정량적인 정확도를 측정하기 위해서 다음과 같은 실험을 수행하였다. 첫 번째로, PARAFAC 분해가 영향력 있는 블로그들을 올바르게 식별했는지를 확인하기 위해서 각 커뮤니티 내에서 높은 점수를 받은 상위 5개의 블로그들이 스크랩한 포스트들을 분석하였다. 일반적으로 많은 블로그들로부터 스크랩을 받는 포스트는 권위 있는 포스트로 간주되며, 권위 있는 많은 포스트들을 스크랩하는 블로그는 영향력 있는 블로그로 간주된다. 실험 결과 높은 점수의 블로그들로부터 스크랩을 받은 포스트들의 평균 스크랩 수가 전체 포스트들의 평균 스크랩 수보다 약 19배 높은 것으로 나타났다. 따라서 PARAFAC 분해가 영향력 있는 블로그들을 올바르게 식별한 것을 확인할 수 있다.

두 번째로, PARAFAC 분해가 권위 있는 포스트들을 올바르게 식별했는지를 확인하기 위해서 각 커뮤니티 내에서 높은 점수를 받은 포스트들을 스크랩하는 블로그들의 수를 분석하였다. 표 3은 도출된 각 커뮤니티 내 포함되어 있는 블로그들 중에서 높은 점수를 받은 상위 k개의 포스트를 스크랩한 블로그들의 평균 비율을 나타낸 것이다. 대부분의 블로그들은 해당 커뮤니티 내에서 높은 점수를 받은 포스트들을 스크랩하였다. 권위 있는 포스트는 많은 블로그들로부터 스크랩을 받기 때문에 PARAFAC 분해가 권위 있는 포스트들을 올바르게 식별한 것을 확인할 수 있다.

요약하면, PARAFAC 분해는 블로그 공간에서 커뮤니티들을 파악해 냈으며, 각 커뮤니티 내에서 영향력 있는 블로그들과 키워드, 그리고 권위 있는 포스트들도 성공적으로 식별하였다.

<표 2> 첫 번째 커뮤니티내의 영향력 있는 키워드들과 권위 있는 포스트들의 주제 (R=50).

키워드 (C1)	상위 블로그에서 출현 빈도가 높은 키워드 (C1)
영어, 법칙, 발음, 미국	영어, 생활, 묘사, 대화
	미국, 발음, 법칙, 영어
	미국, 발음, 법칙, 영어
	미국, 생활, 튀김, 치즈
	상위 포스트 제목 (C1)
	한국인들이 어려워하는 미국 영어 발음법칙 35개
	영어발음공부
	전화 받을 때 유용한 영어 표현들
	유용한 영어표현

<표 3> 권위 높은 상위 k개의 포스트를 스크랩한 블로그들의 비율.

k=1	k=3	k=5	k=10
75.87%	87.53%	93.06%	96.70%

3. 결론

본 논문에서는 블로그 공간을 텐서 형태로 모델링하고, PARAFAC 분해를 이용하여 분석하였다. PARAFAC 분해는 커뮤니티 식별과 각 커뮤니티 내에서 영향력 있는 블로그와 키워드, 그리고 권위 있는 포스트를 찾아내는데 유용한 것으로 나타났다.

감사의 글

본 연구는 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단 (No.2008-0061006) 및 지식경제부 및 정보통신산업진흥원의 'IT융합 고급인력과정 지원사업 (NIPA-2011-C6150-1101-0001)과 NHN(주)의 지원을 받았습니다. 그러나, 본 논문에서 제시된 의견이나 결론, 또는 권고 등은 온전히 저자(들)의 것이며, 반드시 지원회사의 입장을 대변하는 것은 아닙니다.

참고문헌

[1] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading Behavior in Large Blog graphs," In *Proc. of SIAM Int'l. Conf. on Data Mining*, SDM, 2007.

[2] S. Yoon, J. Shin, S. Kim, and S. Park, "Extraction of a Latent Blog Community Based on Subject," In *Proc. of ACM Int'l. Conf. on information and Knowledge Management*, ACM CIKM, pp. 1529-1532, 2009.

[3] T. Kolda, and B. Bader, "Tensor Decompositions and Applications," *Journal of SIAM Review*, Vol. 51, No. 3, pp. 455-500, 2009.

[4] B. Bader and T. Kolda, "Algorithm 862: MATLAB Tensor Classes for Fast Algorithm Prototyping," *ACM Transactions on Mathematical Software*, Vol. 32, No. 4, pp. 635-653, 2006.