문서에서의 용어위치 근접성을 이용한 용어간 관계 추출

김진영 엄재홍[O,*] 장병탁
서울대학교 전기·컴퓨터공학부 바이오지능연구실
myleo.jerry@gmail.com, jheom@bi.snu.ac.kr, btzhang@bi.snu.ac.kr

Extraction of Term-term Relationship using Proximity in Document

Jin-young Kim Jae-Hong Eom[O,*] Byoung-Tak Zhang
Biointelligence Lab., School of Computer Science & Engineering
Seoul National University

연구의 배경

폭발적으로 늘어난 정보를 처리하는 여러 방법들 중에서 '텍스트 기반 지식발견(Text-based Knowledge Discovery)'은 주어진 텍스트 전체에서 여러 종류의 유용한 패턴을 추출하는 기술로, 원시 데이터의 범위를 좁혀 사용자 연관 부분만을 추려내는 검색과는 구별된다. 지식발견과 관련 선도적인 연구로는, 학회지 논문의 수동분석을 통한 Raynaud병의 치료제 발견[1]에서부터 자동화시스템[2], 개방발견법(Open Discovery Process)[3] 등이 제안되어왔다. 연관용어 개념 분석을 통한 지식발견과 함께 분석결과 시각화기법[4,5,6,7]도 제안되어 용어간 관계들에 대한 개념지도(Concept Map)형태를 이용해 새로운 지식을 발견하기도 했다. 지식발견의 대상 중 대표적인 것이 문서에 사용된 용어간의 관계로서, 이를 추출하고 시각화하는 다양한 방법이 제안되어 왔다.

기존 연구의 한계

하지만 기존 연구들은 문서를 용어의 집합으로 간주하기에, 풍부한 정보를 담고 있는 문서 내 용어간의 거리를 무시하는 한계가 있었다. 즉, 문서에 같이 출현하는 단어 사이에 관계가 있다는 것을 전제하고, 문서에 같이 발생(co-occurrence; 공발생)하는 빈도에 따라 관계의 강도를 결정하였지만, 문서 내 용어간의 거리는 고려되지 않았다. 즉, 문서에 같이 출현하는 모든 용어간 관계들의 강도가 동일한 것으로 간주되었다. 하지만 실제 문서에서의 용어간의 거리가 상호 관계에 갖는 의의를 생각하면 이것이 얼마나 많은 정보의 손실을 가져오는지 알 수 있다 (특히, 길이가 길고 다양한 용어를 포함하는 문서의 경우). 기존연구[4,5,7]에서 논문 초록과 같은 일정한 길이를 갖는 동질적인 문서 집합만을 대상으로 실험을 수행한 것은 이와 같은 추출 방식의 한계에 기인한 바가 크다고 할 수 있다.

공발생과 근접성을 고려하는 새로운 용어간 관계 추출법

이러한 관점에서 보면, 문서 내 근접 용어간의 관계에 좀 더 큰 가중치를 두는 방식으로 용어간 근접성을 고려할 경우, 기존의 방법보다 용어 간 관계를 좀 더 정확히 추출할 수 있을 것이다. 이러한 방법은 우선 의미추출의 단위가 세분화되므로 다양한 길이 및 성격을 가지는 문서를 다룰 수 있다. 또한 같은 양의 문서에서 훨씬 풍부한 용어간 관계 정보를 끌어낼 수 있고, 반대로 적은 양의 문서에서도 기존 방법과 같은 수준의 결과를 얻어낼 수 있을 것이다. 이에, 본 연구에서는 용어간의 근접성을 고려한 관계 추출에 연구의 초점을 두었다. 연구에서 우선 주목한 것은 의미 전달 단위로서 단락과 문장의 역할이었다. 즉, 글의 의미는 단락과 문장으로 분절되어 구조화되니, 문서에서 같은 단락의 공발생 단어 간의 관계는 단순히 문서에 함께 나오는 관계보다 강할 것이며, 같은 문장의 공발생 단어 간의 관계는 단락내의 공발생보다 강할 것이라는 것이다. 반면, 어떤 단어 쌍의 개별 단어가 특정 의미 단위에서 단독으로 발생하는 경우가 많이 발견될수록 그들 간의 관계는 약하다고 볼 수 있을 것이다.

예를 들어, 용어 A와 B가 함께 사용된 문서, 단락, 문장의 개수를 각각 $O_D^{AB}, O_P^{AB}, O_S^{AB}$라 하면 두 용어의 근접도 평가치(Proximity Score) $PS_{AB} = W_D \times O_D^{AB} + W_P \times O_P^{AB} + W_S \times O_S^{AB}$와 같이 정의된다. W는 각 문서, 단락, 문장내의 공발생에 대한 가중치상수이다. 두 용어가 각각 따로 쓰인 문서, 단락, 문장의 수를 $O_D^A, O_P^A, O_S^A, O_D^B, O_P^B, O_S^B$라 하면 용어의 단독발생을 고려한 정규화된 근접도 평가치(Normalized Proximity Score) $NPS_{AB} = W_D \times M(O_D^{AB}, O_D^A, O_D^{B)} + W_P \times M(O_P^{AB}, O_P^A, O_P^{B)} + W_D \times M(O_S^{AB}, O_S^A, O_S^{B)}$와 같이 정의되며, M은 용어간의 관계를 나타내는 임의의 측정 지표를 나타낸다. 용어간 관계를 계산하는 여러 방법[8] 가운데 본 논문에서는 가장 널리 사용되는 코사인 유사도(Cosine Similarity) $M_C(a,b,c) = a/\sqrt{(a+b)} \times \sqrt{(a+c)}$를 사용하였다.

문서 내 용어간의 근접성에 대한 정보검색 분야 논문에서 검색 순위에 질의어 단어간 근접성(Query-term Proximity)에 대한 고려를 포함하는 연구결과[9,10]가 있었는데, 이 방법 역시 사용자가 입력한 검색어가 문서에서 가까운 위치에 발견되는 경우가 그렇지 않은 경우보다 더 적절한 결과일 확률이 높다는 점을 이용한 것이다. 이 방식은 우리의 관심사인 용어간 관계 추출에도 적용될 수 있는데, 각 문서에서 발견된 용어의 순서쌍을 위와 같은 방식으로 평가하고, 모든 문서에 대해 평가치를 합산하는 방법을 생각해볼 수 있다. 이를 실제 계산하기 위해서는 각 문서에서 발생한 용어 쌍을 인접한 것끼리 묶는 방법을 결정해야 하는데, 본 연구에서는 서로 겹치지 않으며 각 그룹의 길이(포함 단어 수)가 최소화되게 묶는 방법[11]을 적용하였다.

그룹화 된 용어 쌍의 관계 계산은, 모든 그룹에서 용어 A와 B를 포함한 n개의 그룹 G_k가 발견되고 k번째 그룹의 길이가 l_k, A혹은 B가 그룹에 포함되지 않고 단독으로 발생한 횟수 O_A, O_B에 대해 두 용어의 단어 수준 근접도 평가치(Word-level Proximity Score) $WPS_{AB} = \sum_{k=1}^{n} l_k^{-1} = W_k \times (O_A \times O_B)$ 와 같이 단독 발생에 대한 가중치 상수 W_K를 이용하여 표현 가능하며, WPS_{AB} 값이 음수가 되지 않도록 적절히 결정해야 한다. 이렇게 구한 직접 관계 행렬에서 각 행간의 코사인 유사도 등을 구함으로써 이를 간접 관계로 확장시킬 수 있다. 이렇게 구한 간접적 관계는 비슷한 용어들과 관계를 갖는 두 용어가 서로 높은 관련성을 갖는 것으로 간주하므로, 직접적 관계의 경향을 비교하는 것으로 볼 수 있다. 또한 직접적 관계에서의 수치적 차이가 상관 계수 계산을 통해 상쇄되므로, 자주 발생하는 용어간의 관계가 높게 나타나는 직접적 관계의 문제로부터 자유롭다[12]. 본 연구에서는 근접성 평가치(NPS_{AB})와 단어 수준 근접성 평가치(WPS_{AB})에 대한 간접적 관계를 구하여 비교에 사용하였다.

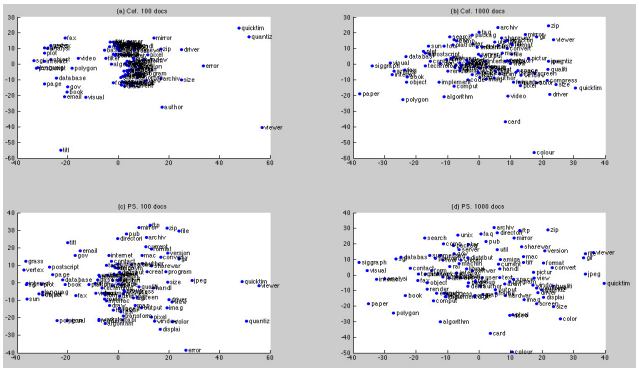

그림 1. comp.graphics 뉴스그룹의 용어지도 (a, b: 공발생만 고려 /c, d: 공발생 및 근접성 고려, a~d: 좌측 위쪽부터 가로 방향으로)

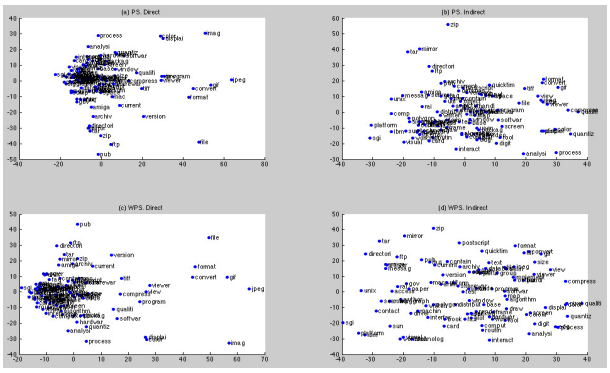

그림 2. comp.graphics 뉴스그룹의 용어 지도 (a, c: 직접적 관계, b, d: 간접적 관계)

실험 결과 및 결론

실험에서는 앞서 살펴본 평가 척도를 이용하여 comp.graphics 뉴스그룹 데이터를 분석하였다. 그림 1은 공발생만 고려한 추출법과 공발생 및 근접성을 고려한 추출법(NPS)을 문서 100개와 1,000개를 대상으로 수행한 후 MDS를 적용한 결과이다. 기존 추출법을 사용한 경우(a ,b) 용어의 군집이 개별 용어를 분간하기 힘든 몇 개의 그룹으로 뭉치는 현상이 나타나지만, 용어간 근접성을 고려한 추출법(c, d)에서는 개별 단어 간의 거리가 적절히 유지되었다. 즉, 추출법이 용어간의 관계를 훨씬 더 많은 단계로 촘촘히 표현하였다. 또한 각각 문서 100개(a, c)와 1,000개(b, d)를 가지고 실험한 결과를 비교하면, 기존 추출법은 문서의 개수에 따라 확연한 성능 차이를 나타내는 반면 개선된 추출법은 적은 수의 문서에서도 만족할만한 결과를 보였다. 그림 2는 근접성 평가치(PS)와 단어 수준 근접성 평가치(WPS)에 대해 직접적 관계(a, c)와 간접적 관계(b, d)를 비교한 것이다. 간접적 관계에서는 용어의 발생 횟수에 따른 편차가 해소되어 용어간 거리가 비교적 균일하게 유지되는 것을 확인할 수 있다.

본 연구에서는 문서에서의 근접성을 고려하여 통계적으로 용어간 관계를 추출하는 방법을 제안하였다. 뉴스그룹 텍스트 실험 결과, 기존 방법에 비해 제안된 방법은 개별 용어의 발생 빈도에 독립적이고 좀 더 세분화된 용어간의 관계를 추출해 내었다.

참고문헌

[1] Swan, D.R., "Fish oil, Raynaud's syndrome and undiscovered public knowledge," *Perspectives in Biology and Medicine*, 30, 7~18, 1986.

[2] Swanson D.R. and Smalheiser N.R., "An interactive system for finding complementary literatures: A stimulus to scientific discovery," *Artificial Intelligence*, 91, 183~203. 1997.

[3] Lindsay, R.K., and Gordon, M.D., "Literature-based discovery by lexical statistics," *Journal of the American Society for Information Science*, 50, 574~587, 1999.

[4] Eijk, CC van der, E.M. van Mulligen, J.A. Kors, B. Mons and J. van den Berg, "Constructing an Associative Concept Space for Literature-based Discovery," *Journal of the American Society for Information Science*, 55, 436~444, 2004.

[5] van Eck, N.J., Waltman, L. and van den Berg, J., "Novel Algorithm for Visualizing Concept Association," *Proceedings of the 16th International Workshop on Database and Expert Systems Applications*, 2005.

[6] Borg, I. and Groenen, P.J.F., "*Modern Multidimensional Scaling*," 2nd Ed. New York: Springer, 2005.

[7] van Eck, N.J. and Waltman, L., "VOS: a new method for visualizing similarities between objects," *Proceedings of the 30th Annual Conference of German Classification Society*, 299~306. Springer, 2007.

[8] Chung, Y.M., and Lee, J.Y, "A corpus-based approach to comparative evaluation of statistical term association measures," *Journal of the American Society for Information Science and Technology*, 52, 283~296. 2001.

[9] Clarke, C.L.A., Cormack, G.V., and Tudhope, E.A., "Relevance ranking for one to three term queries," *Information Processing and Management*, 36, 291~311. 2000.

[10] Song, R., Wen, J., and Ma, W., "Viewing Term Proximity from a Different perspective," *Technical Report MSR-TR-2005-69*, Microsoft Research. 2005.

[11] Justeson, J.S. and Katz, S.M., "Technical terminology: some linguistic properties and an algorithm for identification of terms in text," *Natural Language Engineering*, 9~27, 1995.

[12] McCain, K.W., "Mapping authors in intellectual space: a technical overview," *Journal of the American Society for Information Science*, 41, 433~443, 1990. □

※ 이 논문은 2006년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2006-511-D00355).