하이퍼네트워크 관점에서 본 문서에서의 단어간 긴밀성과 다양성의 대칭성

김준식[01], 박찬훈[2], 장병탁[2]

[1]서울대 물리 천문학과
[2]서울대 컴퓨터 공학과 바이오 지능 연구실
{jskim, chpark, btzhang}@bi.snu.ac.kr

Affinity and Variety between Words in the Framework of Hypernetwork

Joon Shik Kim[01], Chan-Hoon Park[2], Byoung-Tak Zhang[2]
[1]Department of Physics and Astronomy, Seoul National University
[2]Biointelligence laboratory, School of Computer Science and Engineering, Seoul National University

전체 문서에서의 두 단어간 연결 상태를 파악하여 앞 단어 다음에 오는 단어의 빈도수에 따른 그룹 분류로 그 다양성과 긴밀성을 살펴보았다. 기존의 연구에서 Zipf's power law 는 chinese restaurant process 로 설명되었고 scale free network 에서는 edge 수에 따른 노드의 profile 을 조사하여 hub 를 찾는 연구가 수행되었었다. 이처럼 Zipf law 와 그 설명에 관한 논문은 있으나 하이퍼네트워크 관점에서 노드들 간의 연결관계를 설명하는 연구는 없었다. 우리는 이 논문에서 하이퍼네트워크관점에서 정리한 기억의 기본 소자 양식을 이용하여 문서속에서의 단어간 연결을 count 수로 나타내는 방법을 사용한다. 이 방법은 하이퍼네트워크를 기억 인출이나 , 문서 분류의 기능을 수행하는 측면 보다는 기억의 기본 소자로서 이용하여 전체 문서 혹은 기억의 통계적인 분포를 알아보고자 하는데 쓰는 것이다. 우리는 양자역학의 원자 모델에서의 에너지 준위 (energy level)를 유추 (analogy) 했고, 같은 에너지를 가지는 자유도의 갯수인 축퇴도 (degeneracy) 라는 물리량을 사용하여 문서를 분석해 보았다. 본 논문에서는 전체 문서 (corpus) 와 그 일부를 복제한 부분 문서 (query) 를 통해 단어간의 연결에서의 긴밀성과 다양성 간에 대칭적인 성질이 있으며 이는 기계학습의 이용 (exploitation) 과 탐색 (exploration) 으로 설명할 수 있음을 논의 하고자 한다. 데이터 분석 결과는 단어간 연결의 긴밀성과 다양성 간의 대칭성으로 나타나며 이는 exploitation 과 exploration 의 관점에서 설명될 수 있다. 또한 약간의 대칭성 깨짐 (symmetry breaking) 도 TIPSTER data 에서 관찰되었다.

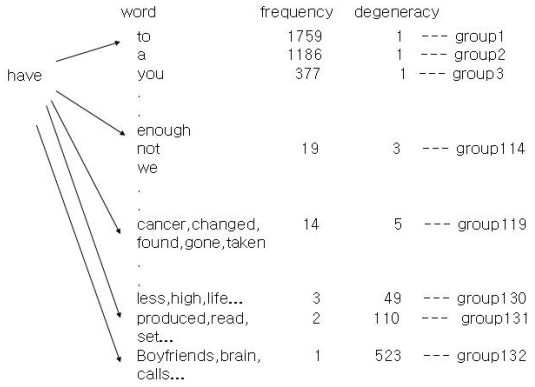

그림 1. 단어의 frequency를 기준으로 분류한 group, frequency 그리고 degeneracy 의 예.

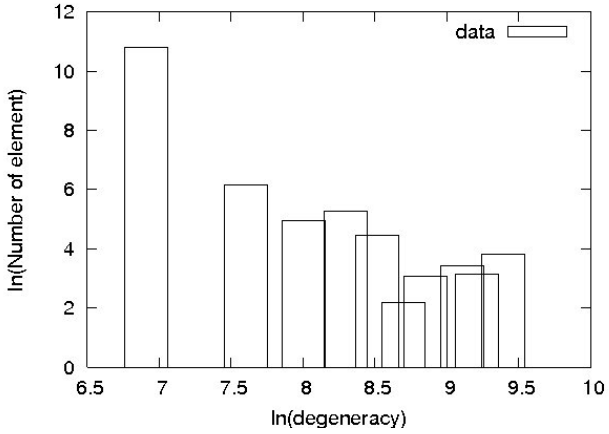

그림 2. data 들을 degeneracy 값들에 따라 돗수 분포표를 그려보았다. x, y 값은 자연로그를 취한 값들이다. 데이터들이 선형적 감소를 보인다. 이 결과는 degeneracy profile 이 power law 를 따름을 보인다.

참고문헌

1. M. Steyvers, T.L. Griffiths, and S. Dennis, Probabilistic inference in human semantic memory, TRENDS in cognitive science 10 (7), 327--334, (2006).
2. P. Bak, K. Christensen, L. Danon, and T. Scanlon, Unified scaling law for earthquakes, Physical Review Letters 88 (17), 178501, (2002).
3. P. Bak and K. Chen, Scale depedent dimension of luminous matter in the universe, Physical Review Letters 86 (19), 4215--4218, (2001).
4. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.-L. Barabasi, Hierarchical organization of modularity in metabolic networks, Science 297, 1551--1555, (2002).
5. C. Furusawa, Zipf's law in gene expression, Physical Review Letters 90 (8), 088102, (2003).
6. A.-L. Barabasi and R. Albert, Emergence of scaling in random networks, Science 286, 509--512, (1999).
7. S. Goldwater, T.L. Griffiths, and M. Johnson, Interpolating between types and tokens by estimating power-law generators, Advances in Neural Information Processing Systems 18, 459--466, (2006).
8. K.E. Kechedzhi, O.V. Usatenko, and V.A. Yampol'skii, Rank distribution of words in correlated symbolic systems and the Zipf law, Physical Review E 72, 046138, (2005).
9. B.-T. Zhang and J.-K Kim, DNA hypernetworks for information storage and retrieval, Lecture Notes in Computer Science, DNA12, 4287, 298--307, (2006).
10. S. Kim, M.-O. Heo, and B.-T.Zhang, Text classifier evolved on a simulated DNA computer, IEEE Congress on Evolutionary Computation (CEC 2006), 9196--9202, (2006).
11. 김준식, 김종찬, 노영균, 이동윤, 장병탁, DNA 컴퓨팅 연산 과정의 통계 물리적 예측, 한국컴퓨터종합학술대회 2005 논문집, 제32권 1(B), 253--355, 2005.07.
12. K.S. Krane, Introductory nuclear physics, John Wiley & Sons, Inc, 1988.
13. S. Maslov, M. Paczuski, and P. Bak, Avanlanches and 1/f noise in evolution and growth models, Physical Review Letters 73 (16), 2162--2165.