상호정보량의 정규화에 대한 연구
A Study on Relative Mutual Information Coefficients

이 재 윤(Jae-Yun Lee)

초 록
상호정보량은 용어간 유사도 산출을 비롯한 다양한 분야에서 연관성 척도로 사용되어왔다. 그러나 값의 범위가 일정하지 않으며 지나치게 저빈도인 경우를 선호하는 경향이 제한점으로 지적되고 있다. 이런 점을 보완하기 위해서 상호정보량을 정규화하는 상대적 상호정보량 계수를 제안하였다. 제안된 계수의 특성을 알아본 다음, 세 실험집단을 대상으로 전역적(global) 질의확장 검색을 수행한 결과 검색 성능을 향상시킬 수 있었다.

ABSTRACTS
Mutual information, as an association measure, has been used for various purposes as well as for calculating term similarity. There are, however, some limits in mutual information. It tends to emphasize low frequency terms extremely because the marginal value of mutual information changes inversely to frequency of terms. To compensate for this limit, this study suggests relative mutual information(RMI) coefficients which normalize mutual information, and examines their characteristics in some details. The RMI coefficients also improve effectiveness of global query expansion when they are adapted to three different collections.

키워드: 상호정보량, 상대적 상호정보량 계수, 연관성척도, 정보검색, 질의확장
Mutual Information, Relative Mutual Information Coefficients, Association Measures, Information Retrieval, Query Expansion

연세대학교 문헌정보학과 강사(memexlee@lis.yonsei.ac.kr)
논문접수일자 2003년 11월 22일
게재확정일자 2003년 12월 11일

참고문헌
강현규. 1997. 『자연언어 정보검색에서 상호정보를 이용한 2단계 문서순위 결정 방법』. 박사학위논문, 한국과학기술원.
김명철, 이운재, 최기선, 김길창. 1992. “시소러스 작성을 위한 개념 획득 도구”. 제4회 한글 및 한국어 정보처리 학술대회 논문집, 39-49.
김성혁, 서은경, 이원규, 김명철, 김영환, 김재군. 1994. “자동색인기 성능시험을 위한 Test Set 개발”. 정보관리학회지, 11(1): 81-102.
김정세. 1996. 『정보 검색에서 상호 정보를 이용한 문서 순위의 재조정』. 석사학위논문, 계명대학교.
김지영, 장동현, 맹성현, 이석훈, 서정현, 김현. 2000. “한국어 테스트 컬렉션 HANTEC의 확장 및 보완”. 제12회 한글 및 한국어 정보치리 학술대회 논문집, 210-215.
김판구. 1994. 『한국어 정보 검색을 위한 상호 정보량에 기반한 복합어 자동 색인』. 박사학위논문, 서울대학교.
이공주, 김재훈, 김길창. 1995. “품사태깅된 말뭉치로부터 한국어 연어 추출”. 95 가을 한국정보과학회 학술발표논문집, 623-626.
이재윤. 2003. “유사계수에 따른 전역적 질의확장 검색 성능 비교”. 2003 한국정보과학회 가을 학술발표논문집(I), 526-528.
이재윤. 2003a. “질문 유형에 따른 인터넷 검색엔진의 성능 비교”. 제10회 한국정보관리학회 학술대회 논문집, 185-192.
장명길. 2002. 『교차언어 정보검색에서 상호정보를 이용한 사전기반 질의변환』. 박사학위논문, 충남대학교.
정석원. 2001. 『격 관계와 상호정보를 이용한 한국어 의존 파서에 관한 연구』. 석사학위논문, 연세대학교.
Abramson, Norman. 1963. Information Theory and Coding. New York: McGraw-Hill Book Company.

Astola, Jaakko, and Ilkka Virtanen. 1982. “Entropy correlation coefficient: a measure of statistical dependence for categorized data”. Proceedings of the University of Vaasa, Discussion Papers No. 44, 1982. Quoted in Maes et al. 1997.
Buckley, Chris, Gerard Salton, and James Allan. 1993. “Automatic retrieval with locality information using SMART”. In D. K. Harman (ed.), Proceedings of the First Text REtrieval Conference (TREC-1), (pp. 59-72). NIST Special Publi- cation 500-207.
Chung, Young Mee, and Jae Yun Lee. 2001. “A corpus-based approach to comparative evaluation of sta-tistical term association measures”. Journal of the American Society for Information Science and Tech-nology, 52(4): 283-296.
Church, K. W., and P. Hanks. 1990. “Word association norms, mutual information, and lexicography”. Computational Linguistics, 16(1): 22-29.
Efthimiadis, E.N. 1996. “Query expansion”. In M.E. Williams (ed.), Annual Review of Information Science and Technology, (v.31) (pp.121-187). Information Today, Inc., Medford, NJ.
Fano, Robert M. 1961. Transmission of Information: A Statistical Theory of Communications. New York: M.I.T. Press.
Kim, M. C., and K. S. Choi. 1999. “A comparison of collocation-based similarity measures in query expansion”. Information Processing & Management, 35(1): 19-30.
Nie, N. H., C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent. 1975. SPSS: Statistical Package for the Social Sciences. 2nd ed. New York: McGraw-Hill Book Company.
Maes, F., A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. 1997. “Multimodality image registration by maximization of mutual infor-mation”. IEEE Transactions on Medical Imaging, 16(2) : 187-198.
Mandala, R., T. Tokunaga, and H. Tanaka. 1998. “Query expansion using heterogeneous thesauri”. Information Processing & Manage-ment, 36(3): 361-378.
Otte, M. 2001. “Elastic registration of fMRI data using Bezier-spline transformations”. IEEE Transactions on Medical Imaging, 20(3): 193-206.
Pluim, J. P. W., J. B. A. Maintz, and M. A. Viergever. 2003. “Mutual-information-based registration of medical images: a survey”. IEEE Transactions on Medical Imaging, 22(8) : 986-1004.
Qiu, Y., and H. P. Frei. 1993. “Concept based query expansion”. Procee-dings of the 16th Annual Inter-national ACM-SIGIR Conference on Research and Development in Information Retrieval, 160-169.
Shannon, C. E. 1948. “A mathematical theory of communication”. Bell System Technical Journal, 27: 379-423, 623-656.
Salton, Gerard. 1968. Automatic Infor-mation Organization and Retrieval. New York : McGraw-Hill Book Company.
Spth, Helmuth. 1980. Cluster Analysis Algorithms for Data Reduction and Classification of Objects. Ellis Horwood Limited.
Strehl, Alexander, and Joydeep Ghosh. 2002. “Cluster ensembles: A knowledge reuse framework for combining multiple partitions”. Journal of Machine Learning Research, 3(Dec) : 583-617.
Studholme, Colin. 1997. Measures of 3D Medical Image Alignment. Ph. D. thesis, University of London.
Yang, Yiming, and J. P. Pedersen. 1997. “A comparative study on feature selection in text categorization”. Proceedings of the Fourteenth International Conference on Ma-chine Learning (ICML’97), 412-420.