동시링크를 이용한 웹 문서 클러스터링 실험*

Clustering of Web Document Exploiting with the Co-link in Hypertext

김 영 기(Young-Gi Kim)**
이 원 희(Won-Hee Lee)***
권 혁 철(Hyuk-Chul Kwon)****

〈목 차〉

초 록

인간은 지식의 조직을 통해 세계를 이해한다. 정보검색분야에서 연구되고 있는 정보의 조직화에는 분류와 클러스터링이라는 두 가지 유형이 있다. 분류는 미리 정의된 범주에 각 항목을 배정하는 행위인 반면, 클러스터링은 유사하거나 관련된 항목을 집단화함으로써 정보를 조직한다. 인터넷 정보자원의 조직은 웹 문서에 출현하는 단어들에서 키워드를 추출하여 역파일을 작성함으로써 검색에 활용하는 것이 일반적인 방법이다. 그러나 키워드의 출현 위치나 단어빈도를 통한 문서유사도 기법은 사용된 언어가 다르거나 대부분이 앵커텍스트만으로 구성되어 있는 대문페이지처럼 적용하기 어려운 경우가 많다. 이 연구는 계량정보학적 분석 기법 중에서 동시인용 기법을 웹 문서의 하이퍼링크에 적용하여, 웹 문서의 클러스터링 가능성을 실험한다.

주제어: 웹 문서 조직, 웹 문서 클러스터링, 동시링크

Abstract

Knowledge organization is the way we humans understand the world. There are two types of information organization mechanisms studied in information retrieval: namely classification and clustering. Classification organizes entities by pigeonholing them into predefined categories, whereas clustering organizes information by grouping similar or related entities together. The system of the Internet information resources extracts a keyword from the words which appear in the web document and draws up a reverse file. Term clustering based on grouping related terms, however, did not prove overly successful and was mostly abandoned in cases of documents used different languages each other or door-way-pages composed of only an anchor text. This study examines infometric analysis and clustering possibility of web documents based on co-link topology of web pages.

Key Words: Web document categorization, web document clustering, co-link

* 이 논문은 과학기술부(한국과학기술기획평가원)의 국가지정연구실 사업지원으로 이루어진 것임.
** 부산대학교 문헌정보학과 강사, 부산대학교 한국어정보처리연구실 IR팀(ykiki6292@hanafos.com)
*** 부산대학교 한국어정보처리연구실(whlee@pusan.ac.kr)
**** 부산대학교 전자전기정보컴퓨터공학부 교수(hckwon@pusan.ac.kr)
• 접수일 : 20032. 5. 20 • 최초심사일 : 2003. 6. 6 • 최종심사일 : 2003. 6. 14

Ⅰ. 서 론

1. 연구목적

인터넷 정보자원의 급격한 증대로 이를 효과적으로 조직하여, 검색성능을 향상시키기 위한 다양한 연구와 실험이 수행되고 있다. 인터넷 정보자원의 조직은 웹 문서에 출현하는 단어들에서 키워드를 추출하여 역파일을 작성함으로써 검색에 활용하는 것이 일반적인 방법이다. 그러나 전통적인 정보검색에서 사용해 왔던 키워드의 출현 위치나 단어빈도를 통한 문서유사도 기법을 웹상의 하이퍼텍스트에 적용하는 데는 다음과 같은 몇 가지 문제가 있다.

① 문서의 구조화 정도가 매우 낮고 불분명한 내용이 많다.
② 사용된 언어가 다를 경우 키워드 유사도 기법 적용이 불가능하다.
③ 대문페이지(door way page)는 매우 중요한 정보원임에도 불구하고 텍스트가 아닌 이미지 위주로 구성되어 있는 경우가 많아 키워드를 추출하기 어렵다.

따라서 테이블과 같은 웹 문서의 구조정보나[1] 하이퍼링크와 같은 외부 정보원의 부가적 이용을 고려해 볼 수 있다. 웹 문서의 하이퍼링크를 분석하기 위해서는 인쇄문헌의 인용 분석 기법을 적용할 수 있을 것이다.

인용사항을 색인표목으로 사용한 색인을 인용색인이라 한다. 인용문헌을 이용하여 문헌간의 주제적 관계를 설정하는 기법으로 서지결합법과 동시인용 기법이 있다. 이 두 기법은 유사 주제의 문헌 집단화 및 특정한 정보 요구에 대한 문헌 검색에 이용된다.[2] 서지결합법은 여러 개의 문헌이 공통으로 하나 이상의 문헌을 인용하고 있을 때 이 문헌들은 서로 주제적으로 관련되어 있다는 가정 하에 제시된 것으로, 이 때 문헌들은 서지적으로 결합되어 있다고 말하며[3], 문헌간의 결합도는 공통으로 인용된 문서수로 측정된다. 따라서 결합도가 높을수록 두 문헌의 주제는 유사하다고 본다. 그리고 동시인용 기법은 두 편의 인용문헌이 후에 출판된 제3의 문헌 속에 동시에 인용되었을 때 이 두 편의 문헌들은 서로 주제적으로 관계가 있다는 가정 하에 이러한 제3의 문헌이 많으면 많을수록 두 문헌간의 주제적 유사도는 높다고 보는 것이다.[4]

---

1) 정성원, 이원희, 김영기, 권혁철, "웹 문서 중 의미 있는 표의 추출", 한글 및 한국어 정보처리, 제14집 (2002, 10), pp.332~339.
2) 국민상, 정영미, "인용문헌을 이용한 검색 성능 향상에 관한 실험적 연구", 제19회 한국정보관리학회 학술대회 논문집 (2002, 8), pp.235-240.
3) M. M. Kessler, "Bibliographic coupling between scientific papers", *American Documentation*. Vol.14 No.1(1963), pp.10~25.
4) H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents", *Journal of American society for Information Science*. Vol.24(1973), pp.265~269.

이 연구는 이와 같은 계량정보학적 분석 기법 중에서 동시인용 기법을 웹 문서의 하이퍼링크에 적용하여, 웹 문서의 클러스터링 가능성을 실험한다. 클러스터링(문헌 분류)은 가장 원시적인 자료관리 행위이면서, 동시에 고도의 지식과 테크닉을 요구하는 일로서[5] 정보자원의 효과적인 탐색과 이용을 위한 출발점이다. 웹 문서는 각종 정보원들이 링크를 통해 서로 연결되어 있는데, 웹 문서의 링크는 우선 문서 내 링크(intra-document link)와 문서 간 링크(inter-document link)로 나눌 수 있으며, 나가는 링크(out-link)와 들어오는 링크(in-link)로로 나눌 수도 있다. 나가는 링크가 많은 사이트는 hubness가 높으며, 들어오는 링크가 많은 사이트는 authority가 높다고 말한다.[6]

이 연구는 authority가 높은 문서를 대상으로 한다. 그것은 인쇄문헌의 계량정보학적 분석에서 인용빈도가 높은 문헌을 대상으로 하는 것과 동일한 원리이다. 이 연구에서 사용된 가정은 다음과 같다.

① 웹 문서 A와 B를 제3의 웹 문서가 동시에 링크하고 있다면 A와 B는 주제적으로 관계가 있으며, 제3의 웹 문서가 많으면 많을수록 A와 B의 주제적 연관도가 높다고 가정한다.

② 웹 문서 A와 B, B와 C 사이에 각각 주제적 연관도가 있으면 A와 C의 연관도도 있다고 가정한다.

③ 이런 방법으로 웹 문서를 주제적 연관도에 따라 군집화 시킬 수 있을 것이며, 이런 기법은 웹 문서 클러스터링, 질의확장, 검색성능 향상에 적용될 수 있을 것이다.

2. 관련연구

인터넷 정보자원을 지능적으로 분류하기 위해서는 기존의 분류체계나 분류 프로파일을 가지고 기계학습(machine learning) 방법에 의해 대상 자료를 적절한 주제범주로 배정하는 방법과 유사한 자료들을 동일한 군집에 속하게 묶어주는 클러스터링 기법을 사용할 수 있다. 대표적인 모델로는 규칙기반 모델(Rule-based Model)과 연역적 학습 모델(Inductive Learning Model), 그리고 검색을 활용한 모델을 들 수 있다. 규칙기반 모델은 학습 문서들에 나타나는 범주간의 구별된 규칙을 전문가가 찾아 주거나 학습을 통해 추출된 규칙을 이용하여 문서를 분류하는 모델이다.[7] 연역적 학습 모델에는 학습 문서에서 자질을 추출하여 이를 확률적인 접근 방법으로 사용한 베이지언(Bayesian) 모델[8], 트리

5) 최정태, 양재한, 도태현, 문헌분류의 이론과 실제(부산 : 부산대학교 출판부, 1998), p.i.

6) R. K. Belew, *Finding Out About : A Cognitive perspective on search engine technology and the WWW.* (Cambridge University Press, 2000), p.196.

7) Chidanand Aptè, Fred Damerau, and Sholom M. Weis, "Towards language independent automated learning of text categorization models", *Proc. of the 17th Annual International ACM-SIGIR* (1994), pp.233~251.

8) L. Douglas Baker and Andrew K. Maccallum, "Distributional clustering of words for text classification", *Proc.*

구조로 표현하여 자질의 유무로 범주를 결정하는 결정 트리(Decision Tree) 모델[9], 학습 문서를 통해 생성된 양성 자질(positive feature)과 음성 자질(negative feature)을 벡터 공간으로 표현하고 이들 차이를 극명하게 하는 벡터인 지원 벡터(support vector)를 찾는 SVM(support Vector Machine)[10]이 있다. 한편 정보검색의 관점에서는 분류할 대상 문서를 질의로 보고 이와 유사한 문서를 찾는 방법인 최근린법(K-nearest Neighbor)[11]과 적합성 피드백(relevance feedback)을 기초로 이를 분류에 응용한 Roccio 모델[12]이 있다. 또한 하이퍼텍스트를 대상으로 MRF(Marcov Random Field)기법을 통해 하이퍼링크를 활용하거나[13], 문서 내의 구조적인 정보를 활용하는 연구[14]가 시도되고 있다.

이와 병행하여 하이퍼텍스트를 주제별로 분류하여 관리하기 위한 분류모델에 관한 연구도 활발히 진행 중이다. 대상 문서와 링크로 연결된 이웃 문서의 내용 및 범주를 분석하여, 이웃 문서에 포함된 용어를 반영함으로써 대상 문서의 내용을 확장 해석하고, 이웃 문서의 가용 분류정보가 있는 경우 이를 참조함으로써 정확도 향상을 기한다. 일반 문서 분류 모델의 경우 문서에 출현하는 용어만을 사용하여 분류하는 반면 링크 기반 분류모델은 하이퍼텍스트 내의 링크 정보를 활용하여 대상문서와 연결된 문서를 참조함으로써 분류의 정확도를 향상시키는 모델이다.[15]

Ⅱ. 연구방법

1. 분석 대상 선정

이 연구에서는 웹 문서의 동시인용 분석 실험 대상으로 통계학 분야를 선정하여 주제

---

of the 21th Annual International ACM-SIGIR, 1998.

9) David L. Lewis and Marc. Ringuette, "A comparison of two learning algorithms for text categorization", *Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval* (1998), pp.96~103.

10) Thorsten Joachims, "Text categorization with support vector machines", *Proc. of European Conference on Machine Learning, ECML '98* (1998), pp.137~142.

11) Leah S. Larkey, "Automatic essay grading using text categorization techniques", *Proc. of the 21th Annual International ACM-SIGIR* (1998), pp.90~95.

12) David L. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka, "Training algorithms for linear text classifier", *Proc. of the 19th Annual International ACM-SIGIR* (1996), pp.298~315.

13) Soumen Chakrabarti, Byron Dom, and Piotr Indyk, "Enhanced hypertext categorization using hyperlinks", *Proc. of International Conference on SIGMOD '98* (1998), pp.307~318.

14) 정상화, 이종혁, "문서구조 정보에 기반한 웹 페이지 범주화 모델", 제10회 한글 및 한국어 정보처리 학술대회 (1998), pp.91~96.

15) 오효정, 임정묵, 이만호, 맹성현, "점진적으로 계산되는 분류정보와 링크정보를 이용한 하이퍼텍스트 문서 분류 모델", 제11회 한글 및 한국어 정보처리 학술대회 (1999), pp.89~96.

분야의 지적 구조를 분석하였다. 통계학 분야는 전통적으로 컴퓨터 과학과 밀접한 관련을 갖고 있기 때문에 웹 문서가 풍부하고 다양한 응용을 갖고 있어 분석 대상으로 적절하다고 판단하였다. 이 분야의 웹 문서를 선정하기 위해 주제별 정보원을 제공하는 사이트(정보원 가이드)를 이용하여 얻어진 자료를 링크 순으로 배열하여, 많은 링크를 포함하고 있는 정보원을 대상으로 분석을 시도하였다.

웹 문서의 동시인용 분석을 위한 실험대상 선정 방법은 다음과 같다. 우선 해당 분야의 전문가가 직접 웹 문서를 선별하고 주제 분류를 하여 이용자에게 안내해 주고 있는 인터넷 학술정보원 가이드(Internet resource guide)인 LII[16], The Argus Clearinghouse[17], INFOMINE(Scholarly Internet Resource Collections) Subject List - Statistics[18], Statistics: The World-Wide Web Virtual Library[19], Galaxy[20] 등을 활용하여 선정 후보 목록 작성하였다. 이 중에서 들어오는 링크 수가 1000개 이상인 웹 사이트 53개를 선정하였다. 들어오는 링크 수는 AltaVista와 Google의 "link:url" 검색을 통해, 그리고 자기인용(문서 내 링크)을 제외한 링크 수는 AltaVista의 "link:url and not url", 또는 Google의 결과 내 재검색을 통해 알 수 있다.

해당 주제전문가들이 웹 문서들을 선별하고 분류해 놓은 정보원 가이드들은 분석 대상 사이트를 성격과 주제에 따라 범주화시키기 위한 작업에도 활용되었다. 정보원 가이드들 간의 분류가 일치한 경우는 해당 범주로, 일치하지 않은 경우는 많이 분류된 범주 둘을 택하여 이중으로 범주화하였다. G1, G4, G5에 속하는 문서들은 대부분 가이드들 간의 분류가 일치하였으며, G2와 G6에는 상이한 것이 많았다. 이를 통해 최종적으로 부여된 하위 주제명과 해당 분야 사이트 수는 다음과 같다.

〈표 1〉 전문가에 의한 하위 범주별 문서 수

그룹명	하위 범주	문서수	
		동일범주	이중범주
G1	교육/학습(Online Educational Resources)	7	3
G2	데이터 소스(Data Sources)	1	13
G3	S/W(Statistical Software)	4	3
G4	레퍼런스(References and Service)	5	2
G5	정부기관(Government Statistical Institutions)	19	1
G6	조직/단체(Statistical Research Groups, Institutes, and Associations)		6
G7	저널(Statistical Journals)	2	2

16) http://lii.org/search/file/statistics
17) http://www.clearinghouse.net
18) http://infomine.ucr.edu/dbase/cache/physci/subjects-S.html
19) http://www.stat.ufl.edu/vlib/statistics.html
20) http://www.galaxy.com/galaxy/Science/Mathematics/Statistics/

〈표 2〉 분석대상 문서목록

번호	정 보 원(url/title)	그룹명	링크수
1	http://davidmlane.com/hyperstat/hyperstat online textbook	G1	3167
2	http://davidmlane.com/hyperstat/index.html HyperStat by David M. Lane at Rice U.	G1	1430
3	http://www.hookup.net/ The Introductory Statistics Course: A New Approach	G1	1997
4	http://www.isds.duke.edu/ Institute of Statistics & Decision Sciences	G1/G2	1837
5	http://www.psychstat.smsu.edu/ Introductory Statistics by David Stockburger	G1	1564
6	http://tile.net/lists/ TILE.NET(Internet Reference-discussion and information)	G4	3437
7	http://www.itools.com/ Best Internet Tools(Desk Reference)	G4	18194
9	http://www.utexas.edu/world/lecture/ World Lecture Hall(교육)	G1	10235
8	http://www.lsoft.com/lists/listref.html catalog of LISTSERV lists	G4	4900
10	http://gams.nist.gov/GAMS: Guide to Available Mathematical Software	G3	3532
11	http://lib.stat.cmu.edu/ StatLib at Carnegie Mellon U(Data, Software and News from the Statistics Community)	G3/G2	10654
12	http://lib.stat.cmu.edu/DASL/ DASL : The Data and Story Library for Teacher(online library of datafiles and stories that illustrate the use of basic statistics methods)	G3	1440
13	http://mathforum.org/ Research and education enterprise of Drexel Univ.	G1/G2	28945
14	http://shazam.econ.ubc.ca/ Coin Flipping (econometrix software)	G3/G2	1270
15	http://sosig.esrc.bris.ac.uk/ Social Science Information Gateway)	G4	3457
16	http://sunsite.univie.ac.at/ SunSITE, Austria(This site is mostly devoted to Computer Support for Mathematics, Statistics & Science Education but we also offer information and software for more general areas.)	G1	1401
17	http://www.aihw.gov.au/Australian Institute of Health and Welfare (Australia's national agency for health and welfare statistics and information)	G5	4191
18	http://www.bls.gov/ U.S. Bureau of Labor Statistics	G4/G2	33229
19	http://www.census.gov/ U.S. Census Bureau	G4/G2	332410
20	http://www.kma.go.kr/ 기상청	G5	11879
21	http://www.math.uah.edu/stat/ Virtual Laboratories in Probability and Statistics (web-based resources for students and teachers of probability and statistics)	G1	1407
22	http://www.netlib.org/ Netlib (Netlib is a collection of mathematical software, papers, and databases)	G3	22380
23	http://members.aol.com/johnp71/javastat.html Perform Statistical Calculations (statistical software package)	G1/G3	1514
24	http://www.gnuplot.info/ Gnuplot(S/W)	G3	1020
25	http://demography.anu.edu.au/VirtualLibrary/Demography and Population Studies (The Internet Guide to Demography and Population Studies)	G4	1275
26	http://www.mathtools.net/Mathtools for Science and Engineering (S/W)	G2	6198
27	http://ejde.math.swt.edu/ Electronic Journal of Differential Equations	G7	1101
28	http://etna.mcs.kent.edu/ Electronic Transactions on Numerical Analysis	G7	1099
29	http://www.amstat.org/publications/jse/ Journal of Statistics Education	G7/G2	1558
30	http://www.wiley.com/ John Wiley & Sons Publishers	G7/G2	236811
31	http://www.abs.gov.au/ Australian Bureau of Statistics	G5	12813
32	http://www.cbs.nl/ Statistics Netherlands	G5	12443
33	http://www.cso.ie/ Irish Central Statistics Office	G5	11283
34	http://www.dst.dk/ Statistics Denmark	G5	3460
35	http://www.ine.es/Instituto Nacional de Estadistica - Spain	G5	7786
36	http://www.ine.gov.bo/ National Statisics Institute -- Bolivia	G5	1272
37	http://www.inegi.gob.mx/ Instituto Nacional de Estadistica, Geografia e Informatica	G5	10671
38	http://www.isical.ac.in/ Indian Statistical Institute	G5	1016
39	http://www.istat.it/Italian National Institute of Statistics (ISTAT)	G5	9592

40	http://www.nso.go.kr/National Statistics Office -- South Korea	G5	6585
41	http://www.pcbs.org/Palestinian Central Bureau of Statistics	G5	1337
42	http://www.singstat.gov.sg/ Statistics Singapore	G5	1848
43	http://www.stat.go.jp/ Japanese Statistics Bureau	G5	9532
44	http://www.stat.gouv.qc.ca/Institut de la statistiques du Quebec	G5	2167
45	http://www.statcan.ca/ Statistics Canada	G5	33105
46	http://www.std.lt/ Statistics Department -- Lithuania	G5	1226
47	http://www.unido.org/ Industrial Statistics - UNIDO	G5	8476
48	http://e-math.ams.org/ American Mathematical Society Home Page	G6/G2	5831
49	http://www.amstat.org/ American Statistical Association (ASA)	G6/G2	12997
50	http://www.cbs.nl/isi/ International Statistical Institute (ISI)	G6/G5	5260
51	http://www.imstat.org/ Institute of Mathematical Statistics at UCLA	G2/G6	1274
52	http://www.itl.nist.gov/div898/ Statistical Engineering Division (statistical consulting to the NIST laboratories and performs statistical research.)	G2/G6	1233
53	http://www.siam.org/ SIAM (Society for Industrial and Applied Mathematics)	G2/G6	15907

2. 동시링크 분석

53개의 웹 문서를 대상으로 각 두 개의 문서를 동시에 링크하고 있는 사이트 수를 조사하기 위해 AltaVista의 "고급 웹 검색"을 이용하였다. 나가는 링크(out link)의 경우 웹 문서의 파싱을 통해 그 개수를 쉽게 구할 수 있지만, 들어오는 링크(in link)는 해당 분야의 전체 코퍼스(corpus)를 직접 갖고 있어야만 구할 수 있다. 인터넷 검색 엔진 중에서 들어오는 링크를 검색해 주는 엔진으로 Google과 AltaVista를 들 수 있는데, 그 중에서도 동시링크를 구하기 위한 연산자 사용이 가능한 검색엔진인 AltaVista를 이용하였다. 사용된 검색식은 다음과 같다.

LINK:url A AND LINK:url B

이러한 방법으로 웹 사이트들의 쌍을 동시에 링크하고 있는 웹 문헌들의 검색 건수를 구하여 동시인용빈도 행렬을 작성하였다.

다음으로 웹 문서들 간의 관련성의 정도, 즉 상대적인 유사성과 비유사성을 나타내기 위해 동시인용 빈도는 새로운 척도로 변형되어야 한다. 일반적으로 키워드를 통한 두 문헌간의 관계를 나타내기 위해서는 단어출현빈도(Term Frequency, TF)와 역문헌빈도(Inverse Document Frequency, IDF)를 이용한 TFIDF(TF × IDF)를 이용한 단어벡터(word vector)와 문서 헤드의 거리비교(distance comparison)가 일반적으로 사용된다. 또한 TFIDF를 변형하여 CCIDF(Common Citation × Inverse Document Frequency) 알고리즘을 고려해 볼 수도 있을 것이다. 그러나 이 실험에서는 전체 코퍼스를 갖고 있지 않기 때문에 계산이 불가능한 CCIDF 대신 웹 문서들 간의 상대적인 유사성과 비유사성을 나

타내기 위해 SPSS의 상관계수를 이용하였다.

상관계수는 통상 r로 표기되며 -1에서 1까지의 값을 갖는다. r의 값이 양일 때는 양의 상관관계를, 그리고 음일 때는 음의 상관관계가 있다는 것을 나타내며, 절대값이 1에 가까울수록 상관이 강하다는 것을 의미한다.

다음으로 대상 문서를 분류하기 위해 군집분석(cluster analysis)을 실시하였다. 군집분석의 다양한 방법 중 여기서는 대상 문서간의 유사도를 기초로 전체를 몇 개의 그룹으로 분할하였으며 이를 덴드로그램(dendrogram)으로 나타내었다. 이어서 군집분석의 결과를 시각적으로 나타내기 위해서 다차원척도법(multidimensional scaling)을 이용하여 각각의 웹 문서를 2차원 평면상에 점으로 표시하였다. 다차원척도법은 적당한 성질과 차원을 갖는 공간에 대상과 피험자의 공간배치를 정하는 방법으로서 데이터 속에 잠재해 있는 패턴(pattern)의 구조를 찾아내고, 그 구조를 소수 차원의 공간에 기하학적으로 표현하는데 적절한 방법이다.

Ⅲ. 웹 문헌의 동시인용 빈도

분석대상이 된 53개의 웹 문서를 각각의 쌍으로 만들어 동시인용빈도 행렬을 작성하였다. 이 중에 동시링크 수가 가장 큰 문서 쌍은 D18(U.S. Bureau of Labor Statistics)과 D19(U.S. Census Bureau)로서 이 둘을 동시에 링크하는 문서 수가 8176개에 이른다. 특히 D19는 다른 문서와 평균 605.55개, 전체 32094개의 동시링크를 갖고 있으며 통계학 내의 거의 모든 문서와 함께 매우 높은 수치로 동시 인용됨으로써 이 분야 최고의 authority 사이트임을 알 수 있다. 다음으로 D32(Statistics Netherlands)가 평균 362.64개, 전체 19220개, 그리고 D45(Statistics Canada)가 평균 308.47개, 전체 16349개의 동시링크를 갖고 있는 것으로 나타났다. 하위 범주 중에서는 각 국가의 정부기관(Government Statistical Institutions; D31-D47)의 동시인용 빈도가 가장 높게 나타나고 있다. 이 중에 동시링크 수가 1000개 이상인 문서 쌍을 보이면 다음과 같다.

〈표 3〉 동시링크 수가 1000개 이상인 문서 쌍

번호	문서 쌍	동시링크 수	번호	문서 쌍	동시링크 수	번호	문서 쌍	동시링크 수
1	D18, D19	8176	2	D32, D50	5259	3	D19, D45	3309
4	D 7, D19	1894	5	D20, D40	1797	6	D29, D49	1558
7	D11, D49	1470	8	D11, D12	1470	9	D 1, D 2	1430
10	D45, D53	1362	11	D10, D22	1295	12	D19, D32	1249
13	D22, D53	1143	14	D49, D53	1136	15	D19, D31	1097
16	D19, D43	1076	17	D32, D45	1032	18	D11, D19	1024

여기서 1, 2, 5, 6, 7, 8, 9, 11, 13, 17번 쌍의 경우 모두 같은 하위 범주에 속하는 문서들이지만 나머지 문서 쌍들의 경우는 서로 다른 하위 범주에 속해 있는 문서 쌍들이다. 따라서 단순히 두 문서를 동시에 링크하고 있는 제3의 문서가 많다는 것만으로는 두 문서간의 유사도를 판정하기 어려우며, 특정 범주 내에서 문서간의 상대적인 유사도를 판별할 수 있는 TFIDF나 CCIDF 등과 같은 새로운 척도가 필요하게 된다.

Ⅳ. 동시링크 상관계수 행렬

동시인용빈도를 웹 문서들 간의 관련성의 정도, 즉 상대적인 유사성과 비유사성을 나타내기 위해 통계 프로그램인 SPSS Win 10.0을 사용하여 상관계수 행렬로 변환시켰다. 상관계수가 크면 클수록, 즉 1에 가까울수록 두 문헌의 유사도는 높다. 상관계수는 다음의 과정을 통해 산출할 수 있다.

(1) x의 편차제곱의 합 S(xx)를 계산한다.

$$S(xx)=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2/n$$

(2) y의 편차제곱의 합 S(yy)를 계산한다.

$$S(yy)=\sum_{i=1}^{n}(y_i-\bar{y}=\sum_{i=1}^{n}y_i^2-(\sum_{i=1}^{n}y_i)^2/n$$

(3) x와 y의 편차 곱의 합 S(xy)를 계산한다.

$$S(xy)=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)/n$$

(4) 상관계수 r(Pearson's Correlation)을 계산한다.

$$r=\frac{S(xy)}{\sqrt{S(xx)S(yy)}}$$

일반적으로 상관계수와 그 관련 정도는 다음 표를 기준으로 판단한다.

〈표 4〉 상관계수와 관련정도

상관계수	관련정도
1.0~0.7(-1.0~-0.7)	매우 강한 관련성
0.69~0.4(-0.69~-0.4)	상당한 관련성
0.39~0.2(-0.39~-0.2)	약간의 관련성
0.19~0.0(-0.19~-0.0)	관련성 거의 없음

상관계수 행렬에 따르면 D1과 D2가 상관계수 0.933으로서 가장 높게 나타났다. 그것은

D1과 D2가 실질적으로는 같은 문서이며 서로 다른 이름으로 여러 곳에서 링크를 받고 있기 때문이다. 다음으로 D32와 D50, D11과 D12, D29와 D49의 문서 쌍이 상관계수 0.7 이상으로 매우 강한 관련성을 갖고 있는 것으로 나타났다. 여기서 D32와 D50은 네덜란드의 국가기관과 단체이며, D11과 D12는 둘 다 Carnegie Mellon 대학의 통계관련 자료이고, D29와 D49는 미국 통계협회와 그 학회지이다.

이러한 상관계수를 토대로 대상 문서를 분류하기 위해 근접성 행렬을 작성하여 군집분석(cluster analysis)을 실시하였다. 군집분석의 다양한 방법 중 여기서는 대상 문서간의 유사도를 기초로 전체를 몇 개의 그룹으로 분할하였으며 이를 군집화 일정표와 덴드로그램(dendrogram)으로 나타내었다. 군집화 일정표에 따르면 각 문서는 D1과 D2, D32와 D50이 최초로 묶이며, 이어서 D11과 D12, D29와 D49, D27과 D28의 순서로 군집화 되고 있다. 최종적으로 모두 다섯 개의 군집이 형성되었으며, 그 결과는 다음과 같다. 이 중 G6과 G7은 다른 군집으로 흡수되어 새로운 군집을 형성하지 않았다.

〈표 5〉 동시링크에 의한 군집형성과 문서 수

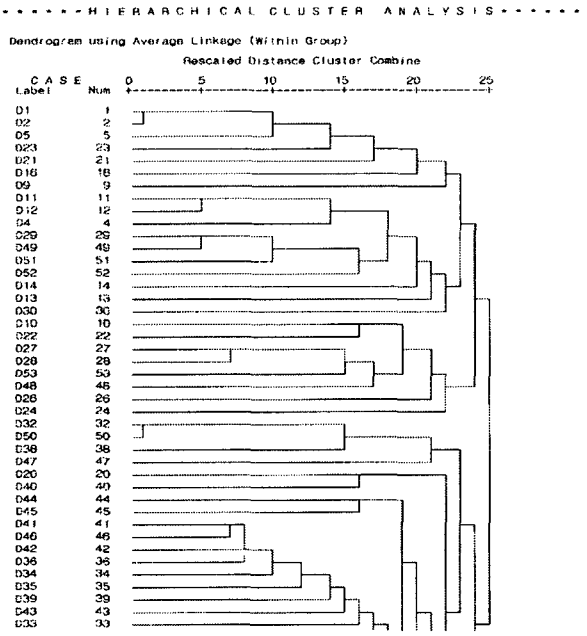
	군 집 명	문서수
C1	교육/학습(Online Educational Resources)	7
C2	데이터 소스(Data Sources)	10
C3	S/W(Statistical Software), 저널(Statistical Journals)	8
C4	레퍼런스(References and Service)	6
C5	기관 및 단체(Government Institutions and other Organization)	22

이어서 웹 문서간의 관계를 시각적으로 나타내어 분석하기 위하여 다차원척도법(multidimensional scaling)을 이용하여 각각의 웹 문서를 2차원 평면상에 점으로 표시하였다. 다차원척도법은 적당한 성질과 차원을 갖는 공간에 대상과 피험자의 공간배치를 정하는 방법으로서 데이터 속에 잠재해 있는 패턴(pattern)의 구조를 찾아내고, 그 구조를 소수 차원의 공간에 기하학적으로 표현하는데 적절한 방법이다.

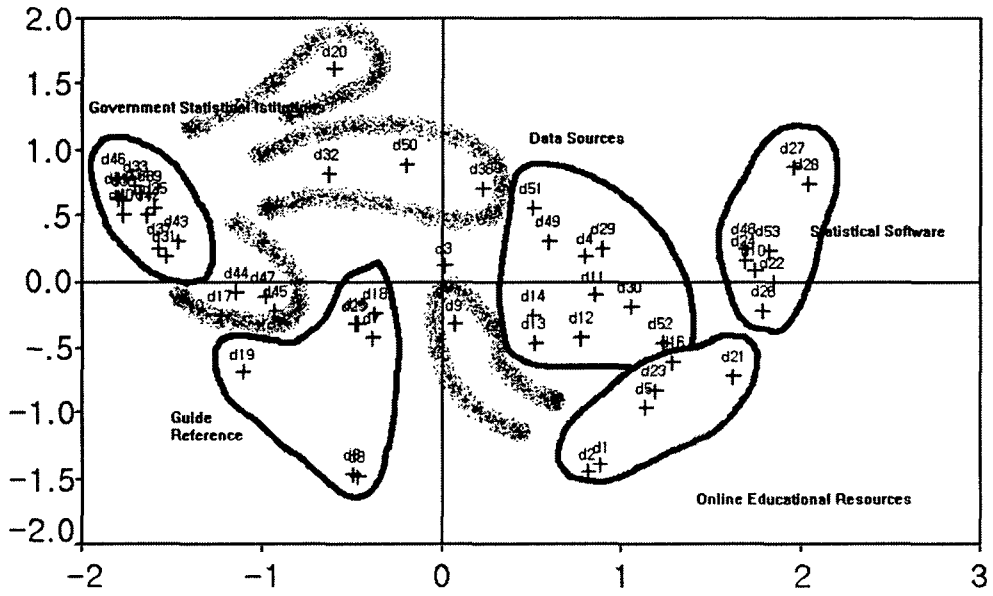
〈표 6〉 군집화 일정표

군집화 일정표

단계	조합된 군집		계수	최초출현 군집단계		다음 단계
	군집 1	군집 2		군집 1	군집 2	
1	1	2	.933	0	0	9
2	32	50	.914	0	0	19
3	11	12	.781	0	0	14
4	29	49	.773	0	0	10
5	27	28	.703	0	0	17
6	41	46	.681	0	0	7
7	41	42	.663	6	0	8
8	36	41	.644	0	7	11
9	1	5	.595	1	0	16
10	29	51	.575	4	0	24
11	34	36	.575	0	8	12
12	34	35	.499	11	0	15
13	19	25	.450	0	0	30
14	4	11	.445	0	3	29
15	34	39	.435	12	0	18
16	1	23	.434	9	0	25
17	27	53	.392	5	0	27
18	34	43	.385	15	0	23
19	32	38	.383	2	0	37
20	10	22	.371	0	0	32
21	20	40	.360	0	0	43
22	44	45	.358	0	0	31
23	33	34	.345	0	18	26
24	29	52	.340	10	0	29
25	1	21	.325	16	0	34
26	33	37	.313	23	0	28
27	27	48	.305	17	0	32
28	31	33	.281	0	26	31
29	4	29	.279	14	24	33
30	18	19	.272	0	13	36
31	31	44	.234	28	22	35
32	10	27	.225	20	27	41
33	4	14	.222	29	0	39
34	1	16	.214	25	0	42
35	17	31	.199	0	31	40
36	15	18	.180	0	30	49
37	32	47	.179	19	0	47
38	6	8	.176	0	0	49
39	4	13	.169	33	0	44
40	7	17	.168	0	35	43
41	10	26	.164	32	0	46
42	1	9	.147	34	0	48
43	7	20	.143	40	21	45
44	4	30	.131	39	0	48
45	3	7	.122	0	43	47
46	10	24	.119	41	0	51
47	3	32	.093	45	37	50
48	1	4	.089	42	44	51
49	6	15	.077	38	36	50
50	3	6	.059	47	49	52
51	1	10	.044	48	46	52
52	1	3	.001	51	50	0



〈그림 1〉 군집화 덴드로그램



〈그림 2〉 다차원척도법에 의한 웹 문서 지도

V. 결론 - 클러스터링 성능 분석

동시링크를 통한 클러스터링 성능을 평가하기 위해 앞에서 살펴 본 것과 같은 전문가들에 의해 분류된 정보원 가이드의 하위범주와 비교하여 정리하면 다음 표와 같다.

〈표 7〉 전문가에 의한 분류와 동시링크를 이용한 클러스터링 실험 결과 비교

전문가들에 의한 분류			동시링크를 통한 클러스터링	
그룹명	문서번호		클러스터명	문서번호
	동일범주	이중범주		
G1	1,2,3,5,9,16,21	4,13,23	C1	1,2,5,21,23,16,9
G2	12	4,11,13,14,18,19,29,30,48,49,51,52,53	C2	11,12,4,29,49,51,52,14,13,30
G3	10,22,24,26	11,14,23	C3	10,22,27,28,53,48,26,24
G4	6,7,8,15,25	18,19	C4	19,25,18,15,6,8
G5	17,20,31-47	50	C5	32,50,38,47,20,40,44,45,41,46,42,36,34,35,39,43,33,37,31,17,7,3
G6		48,49,50,51,52,53		
G7	27,28	29,30		

우선 전문가들에 의한 분류결과가 모두 일치하는 동일범주에 속하는 문서의 경우 G1의 D3과 G4의 D7이 C5로 클러스터링 된 것을 제외하면 동일 범주로 분류되어 있던 문서들은 모두 각각 같은 클러스터로 클러스터링 되었다. 즉 48개의 문서 중 46개가 전문가의 분류와 동시링크를 통한 클러스터링 결과가 일치했다. D3과 D7은 둘 다 일종의 인터넷 검색 툴로서 통계학과는 직접적인 관련이 없는 사이트이며, 덴드로그램에서도 최종 단계에서 C5로 클러스터링 되었다.

통계관련 조직/단체 그룹인G6은 모두 G2(데이터 소스) 또는 G5(정부기관)로 이중 분류되어 있는데, G2에 이중 분류되어 있는 D48-D53 중 D49, D51, D52는 C2로, D48과 D53은 C3으로 클러스터링 되었다. 또한 G7(저널)의 D29와 D30은 G2(데이터 소스)로도 이중 분류되어 있는데, 동일 범주로 분류된 D27과 D28이 C3(소프트웨어)으로 클러스터링 된데 반해 이들은 C2로 클러스터링 되었다.

한편 정보원 가이드에 따라 분류 결과가 서로 다른 이중 범주에 속하는 문서의 경우 역시 D48과 D53을 제외하면 모두 전문가가 분류한 둘 중의 어느 한 범주로 클러스터링 되었다.

전체적으로 보면 전체 53개의 문서를 동시링크 기법을 통해 클러스터링 해 본 결과 4개를 제외한 49개의 문서가 전문가들에 의한 분류와 일치하여 92.5%라는 매우 높은 클러스터링 성능을 보여 주었다. 이 외에도 이 실험을 통해 얻은 성과로 다음과 같은 점을 들 수 있다.

첫째, 이 실험은 통계학이라는 특정 분야 내의 문서를 대상으로 한 단계 더 세분된 범주에서 수행되었기 때문에, 동시링크 기법은 한정된 주제 영역 내의 하위범주를 대상으로 클러스터링 하는 데도 유효한 기법임이 증명되었다.

둘째, 이 실험에는 한국어나 일본어 등과 같은 영어 이외의 언어로 되어 있는 문서도 일부 포함되어 있는데, 이처럼 서로 다른 언어로 되어 있는 문서 간에도 동시링크 기법을 통해 클러스터링이 가능함을 보여 주었다.

셋째, 대문 페이지(door way page)처럼 본문 텍스트가 거의 없이 이미지 위주로 구성되어 있는 경우가 많아 키워드를 통한 문서 유사도 기법의 적용이 어렵지만, 동시링크 기법을 통해서는 클러스터링이 가능하였다.

넷째, 이 실험은 들어오는 링크가 1000개 이상인 문서만을 대상으로 실험이 수행되었는데, 동시링크 기법은 들어오는 링크가 많을수록, 즉 Authority가 높은 문서일수록 클러스터링 성능이 향상됨을 유추할 수 있다.

마지막으로 이 실험의 한계는 다음과 같다.

첫째는 실험대상 문서 수가 웹 전체 문서에 비해 너무 작다는 것이다. 또한 전체 코퍼스를 갖고 있다고 하더라도 클러스터링 결과를 비교 평가할 수 있는 기준을 구하는 것도 어려운 문제이다.

둘째는 통계 프로그램이 제공해 주는 상관계수로는 두 문서간의 상대적인 유사도와 비유사도를 계산하는데 한계가 있다.

셋째는 통계학 이외의 분야에 대한 적용 가능성과 들어오는 링크 수가 작은 웹 문서에도 이 기법이 적용 가능한가 하는 점이 문제로 남아 있다.

또한 동시링크를 이용한 웹 문서 클러스터링 기법과 기존의 클러스터링 기법간의 성능 비교 및 결합, 두 문서간의 상대적인 유사도와 비유사도를 계산하기 위한 보다 적합한 알고리즘의 개발 등에 관한 연구가 후속 과제로 남아 있다.

참고문헌

국민상, 정영미. "인용문헌을 이용한 검색 성능 향상에 관한 실험적 연구", 제19회 한국정보관리학회 학술대회 논문집(2002, 8), pp.235~240.

오효정, 임정묵, 이만호, 맹성현. "점진적으로 계산되는 분류정보와 링크정보를 이용한 하이퍼텍스트 문서 분류 모델". 제11회 한글 및 한국어 정보처리 학술대회(1999), pp.89~96.

정상화, 이종혁, "문서구조 정보에 기반한 웹 페이지 범주화 모델". 제10회 한글 및 한국

어 정보처리 학술대회(1998), pp.91~96.

정성원, 이원희, 김영기, 권혁철. "웹 문서 중 의미 있는 표의 추출". 한글 및 한국어 정보처리 제14집(2002, 10), pp.332~339.

최정태, 양재한, 도태현. 문헌분류의 이론과 실제. 부산대학교 출판부, 1998.

Apté, Chidanand and Damerau, Fred and Weis, Sholom M. "Towards language independent automated learning of text categorization models". *Proc. of the 17th Annual International ACM-SIGIR* (1994), pp.233~251.

Baker, L. Douglas and Maccallum, Andrew K. "Distributional clustering of words for text classification". *Proc. of the 21th Annual International ACM-SIGIR*, 1998.

Belew, R. K. Finding Out About: *A Cognitive perspective on search engine technology and the WWW.* Cambridge University Press, 2000.

Chakrabarti, Soumen and Dom, Byron and Piotr Indyk, "Enhanced hypertext categorization using hyperlinks". *Proc. of International Conference on SIGMOD '98* (1998), pp.307~318.

Joachims, Thorsten. "Text categorization with support vector machines". *Proc. of European Conference on Machine Learning, ECML '98*, 1998. pp.137~142.

Kessler, M. M. "Bibliographic coupling between scientific papers". *American Documentation.* Vol.14 No.1(1963), pp.10~25.

Larkey, Leah S. "Automatic essay grading using text categorization techniques". *Proc. of the 21th Annual International ACM-SIGIR* (1998), pp.90~95.

Lewis, David L. and Ringuette, Marc. " A comparison of two learning algorithms for text categorization". *Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval*(1998), pp.96~103.

Lewis, David L. and Schapire, Robert E. and Callan, James P. and Papka. Ron "Training algorithms for linear text classifier". *Proc. of the 19th Annual International ACM-SIGIR*(1996), pp.298~315.

Small, H. "Co-citation in the scientific literature: A new measure of the relationship between two documents". *Journal of American society for Information Science.* Vol.24(1973), pp.265~269.

부록 1. 통계학분야 웹 문서 동시인용 빈도 행렬

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18
D1	3167	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D2	1430	1430	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D3	0	0	2037	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D4	61	45	9	1837	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D5	462	309	5	73	1564	.	.	.	.	.	.	.	.	.	.	.	.	.
D6	1	0	2	3	1	3437	.	.	.	.	.	.	.	.	.	.	.	.
D7	34	29	5	5	15	119	18194	.	.	.	.	.	.	.	.	.	.	.
D8	7	3	1	1	4	404	183	4900	.	.	.	.	.	.	.	.	.	.
D9	42	38	9	29	46	73	216	65	10235	.	.	.	.	.	.	.	.	.
D10	32	14	5	78	39	7	25	6	45	3532	.	.	.	.	.	.	.	.
D11	336	235	16	556	299	15	33	14	95	538	10654	.	.	.	.	.	.	.
D12	162	125	3	89	162	1	18	3	33	44	1470	1470	.	.	.	.	.	.
D13	73	37	0	24	42	22	189	14	135	123	173	103	28945	.	.	.	.	.
D14	38	15	5	32	42	5	19	1	31	58	188	49	75	1270	.	.	.	.
D15	6	6	6	7	8	18	68	18	102	17	65	19	6	23	3547	.	.	.
D16	22	13	0	30	42	4	4	5	14	29	92	39	41	9	5	1401	.	.
D17	3	1	0	8	6	1	2	4	2	5	11	4	1	0	8	4	4191	.
D18	50	29	2	38	40	27	296	25	100	32	202	59	195	36	60	9	21	33225
D19	242	164	48	148	135	138	1894	147	533	177	1024	399	473	132	603	42	116	8176
D20	0	0	2	5	2	1	4	2	9	2	6	2	0	2	0	0	1	0
D21	173	114	0	58	121	1	2	2	28	33	183	100	69	34	3	24	0	27
D22	49	29	9	86	30	9	35	11	86	1295	743	49	127	56	9	43	7	32
D23	310	199	0	92	165	5	19	15	57	49	270	91	57	40	6	20	6	32
D24	3	2	0	2	0	0	0	0	0	11	14	0	3	0	2	1	0	3
D25	15	9	0	6	7	0	2	3	9	3	40	5	2	0	65	4	24	74
D26	26	17	0	28	20	3	15	4	26	175	152	7	94	16	1	46	2	25
D27	1	1	4	7	2	1	1	0	12	169	45	3	30	3	0	0	0	3
D28	2	1	3	8	4	1	1	1	9	233	69	12	21	4	0	0	0	8
D29	124	82	0	89	96	1	3	1	32	33	390	189	39	19	9	19	6	54
D30	31	24	8	40	26	6	46	21	149	111	170	36	64	35	19	8	3	65
D31	28	16	0	24	26	8	46	9	16	9	63	17	21	6	28	6	674	167
D32	66	43	5	152	79	3	27	7	32	64	434	90	28	34	64	25	38	192
D33	13	5	0	16	18	0	5	1	3	6	42	6	2	0	18	4	26	98
D34	2	1	0	16	3	0	17	1	0	5	28	1	5	0	25	2	18	78
D35	16	6	0	33	13	6	7	1	4	15	69	6	7	3	58	4	21	131
D36	4	0	0	3	5	0	2	0	0	3	17	0	1	0	3	2	15	51
D37	3	0	0	5	3	0	5	0	2	1	37	1	4	0	28	1	16	120
D38	17	10	0	24	6	0	1	0	1	7	68	2	13	0	0	2	9	7
D39	18	8	0	29	22	2	16	1	0	10	64	9	2	1	28	5	30	119
D40	14	4	7	28	17	2	7	4	17	24	68	14	2	18	13	3	32	98
D41	4	0	0	4	4	4	3	0	0	1	18	3	0	0	3	1	12	58
D42	17	7	0	6	9	1	3	0	3	5	37	3	1	7	8	2	24	108
D43	16	6	0	22	13	3	19	4	14	26	88	12	4	22	42	4	30	60
D44	9	7	0	9	6	3	3	2	3	7	33	3	1	1	4	2	16	38
D45	80	59	21	89	62	34	178	37	84	38	308	74	98	62	173	20	65	616
D46	4	0	0	3	6	0	2	0	0	3	9	0	0	0	4	2	15	63
D47	5	0	0	6	4	1	4	1	0	4	40	1	1	4	32	2	9	74
D48	18	9	7	48	17	18	42	7	85	378	268	43	225	17	6	13	0	29
D49	290	192	7	391	231	6	30	13	91	183	1470	374	207	83	59	45	22	280
D50	63	41	5	131	69	2	7	2	17	50	375	76	23	29	15	9	14	65
D51	31	21	0	115	35	2	2	3	16	33	324	36	26	18	4	4	4	29
D52	93	55	0	57	87	2	5	1	14	69	244	89	20	30	1	24	5	29
D53	38	26	11	111	43	15	28	13	102	552	428	64	291	20	6	21	2	87

	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30	D31	D32	D33	D34	D35	D36
D19	33240 8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D20	39	12121	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D21	81	0	1407	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D22	175	5	34	22380	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D23	174	0	114	47	1514	.	.	.	.	.	.	.	.	.	.	.	.	.
D24	6	0	0	65	0	1276	.	.	.	.	.	.	.	.	.	.	.	.
D25	676	0	2	4	4	1	1275	.	.	.	.	.	.	.	.	.	.	.
D26	58	3	26	397	27	16	2	6197	.	.	.	.	.	.	.	.	.	.
D27	13	2	2	108	4	0	1	13	1147	.	.	.	.	.	.	.	.	.
D28	9	2	2	223	2	1	2	23	470	1145	.	.	.	.	.	.	.	.
D29	199	0	57	29	106	0	15	18	31	35	1558	.	.	.	.	.	.	.
D30	237	9	21	159	27	3	2	51	33	49	26	23680 8	.	.	.	.	.	.
D31	1097	1	3	22	23	1	68	8	1	2	34	13	12817	.	.	.	.	.
D32	1249	2	40	65	60	0	40	15	7	18	152	76	595	12443	.	.	.	.
D33	599	0	3	5	4	0	21	3	0	0	23	0	453	826	11285	.	.	.
D34	641	0	0	1	0	0	20	4	0	0	9	2	425	841	660	3460	.	.
D35	868	0	5	6	19	0	28	7	1	1	26	17	508	976	731	736	7785	.
D36	458	0	0	2	2	0	5	2	0	0	6	0	323	368	344	290	373	1272
D37	986	0	0	1	1	0	30	2	0	0	13	2	473	599	456	455	607	521
D38	29	0	2	6	7	0	2	6	1	2	29	7	16	117	14	9	8	8
D39	727	0	4	21	4	0	23	7	1	1	27	5	502	944	735	720	903	341
D40	771	1797	2	13	4	0	13	3	3	3	15	9	482	541	449	387	452	324
D41	463	0	0	2	1	0	19	2	0	0	7	0	320	403	382	324	383	304
D42	752	0	2	3	3	0	23	3	0	0	13	5	524	592	416	422	492	291
D43	1076	4	4	14	8	0	30	6	0	1	17	18	569	690	491	474	568	330
D44	264	0	0	6	4	0	7	4	0	0	8	0	101	119	99	100	105	92
D45	3309	0	24	32	59	1	109	17	3	0	103	59	879	1032	581	599	725	444
D46	433	0	0	2	2	0	7	2	0	0	7	0	360	537	510	468	518	312
D47	324	18	0	5	2	0	29	3	0	0	6	8	79	116	61	65	66	63
D48	147	7	34	350	31	0	2	30	169	186	25	137	3	55	0	4	24	0
D49	872	4	154	188	259	1	33	56	66	76	1558	221	83	850	50	33	78	32
D50	377	2	36	61	56	0	8	14	7	18	142	66	121	5259	176	163	225	109
D51	122	0	35	49	41	1	1	15	4	10	102	71	22	370	22	9	25	8
D52	100	0	74	76	68	6	6	44	3	7	83	25	21	46	14	7	7	6
D53	172	5	45	1143	62	9	2	140	235	369	79	518	12	237	7	3	18	2

	D37	D38	D39	D40	D41	D42	D43	D44	D45	D46	D47	D48	D49	D50	D51	D52	D53
D37	10672	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D38	6	1016	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D39	515	9	9592	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D40	451	16	457	6585	.	.	.	.	.	.	.	.	.	.	.	.	.
D41	372	5	363	448	1337	.	.	.	.	.	.	.	.	.	.	.	.
D42	468	16	445	607	405	1848	.	.	.	.	.	.	.	.	.	.	.
D43	501	16	553	663	410	652	9532	.	.	.	.	.	.	.	.	.	.
D44	143	14	109	96	70	85	101	2167	.	.	.	.	.	.	.	.	.
D45	863	49	697	554	391	606	766	831	33014	.	.	.	.	.	.	.	.
D46	395	6	494	398	333	363	399	83	433	1225	.	.	.	.	.	.	.
D47	80	7	114	90	50	78	91	32	132	53	8476	.	.	.	.	.	.
D48	1	7	8	8	1	0	5	1	25	0	7	5832	.	.	.	.	.
D49	49	135	82	98	20	49	88	36	388	22	31	298	12970	.	.	.	.
D50	141	115	218	174	101	155	182	61	300	129	65	54	800	5257	.	.	.
D51	15	94	24	30	7	24	22	15	97	8	6	63	785	363	1274	.	.
D52	5	8	17	7	5	6	14	5	35	5	4	6	166	40	29	1233	.
D53	1	66	6	16	1	12	20	10	108	3	7	1362	1136	231	287	31	15907

부록 2. 통계학 분야 웹 문서 상관계수 행렬

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
D1	1														
D2	.933(**)	1													
D3	-0.044	-0.047	1												
D4	0.024	0.032	-0.037	1											
D5	.418(**)	.435(**)	-0.047	0.046	1										
D6	-0.05	-0.054	-0.024	-0.053	-0.059	1									
D7	-0.031	-0.029	-0.02	-0.044	-0.041	0.018	1								
D8	-0.045	-0.049	-0.024	-0.053	-0.054	0.176	0.024	1							
D9	-0.028	-0.024	-0.019	-0.029	-0.019	0.002	0.011	-0.006	1						
D10	-0.042	-0.048	-0.038	0.044	-0.022	-0.045	-0.036	-0.045	-0.026	1					
D11	0.088	0.106	-0.031	.314(*)	0.175	-0.04	-0.032	-0.04	-0.022	0.157	1				
D12	0.143	0.167	-0.043	0.24	0.239	-0.053	-0.029	-0.052	-0.024	0.059	.781(**)	1			
D13	-0.02	-0.026	-0.023	-0.033	-0.024	-0.019	-0.006	-0.022	-0.007	-0.004	-0.018	0.006	1		
D14	0.002	-0.004	-0.031	0.032	0.03	-0.037	-0.017	-0.04	-0.009	0.03	0.129	0.109	0.024	1	
D15	-0.041	-0.042	-0.022	-0.037	-0.044	-0.017	0.008	-0.019	0.014	-0.04	-0.012	-0.01	-0.026	-0.006	1
D16	-0.019	-0.021	-0.031	0.004	0.012	-0.031	-0.028	-0.03	-0.021	-0.007	0.035	0.032	0	-0.02	-0.03
D17	-0.05	-0.055	-0.027	-0.049	-0.055	-0.03	-0.026	-0.028	-0.029	-0.05	-0.043	-0.055	-0.027	-0.045	-0.026
D18	-0.022	-0.024	-0.021	-0.019	-0.019	-0.014	0.02	-0.018	-0.006	-0.034	-0.003	0.015	-0.012	0.006	0.023
D19	0.027	0.035	0.001	0.03	0.03	0.015	0.087	0.006	0.029	0.003	0.058	0.144	-0.005	0.066	0.144
D20	-0.049	-0.054	-0.023	-0.05	-0.057	-0.028	-0.026	-0.026	-0.026	-0.047	-0.044	-0.057	-0.025	-0.041	-0.032
D21	0.16	0.165	-0.043	0.061	0.165	-0.049	-0.042	-0.047	-0.021	-0.011	0.114	0.15	0.009	0.022	-0.043
D22	-0.031	-0.034	-0.021	0.008	-0.033	-0.027	-0.025	-0.026	-0.016	.371(**)	0.062	-0.009	-0.015	0.01	-0.031
D23	.286(*)	.293(*)	-0.05	0.112	0.239	-0.052	-0.033	-0.043	-0.007	-0.001	0.168	0.182	-0.011	0.033	-0.039
D24	-0.042	-0.046	-0.023	-0.044	-0.053	-0.027	-0.025	-0.026	-0.026	-0.015	-0.028	-0.047	-0.021	-0.038	-0.029
D25	-0.023	-0.024	-0.029	-0.025	-0.031	-0.026	0.009	-0.027	-0.012	-0.049	0.011	0.031	-0.031	-0.011	0.097
D26	-0.035	-0.038	-0.026	-0.024	-0.037	-0.029	-0.025	-0.027	-0.021	0.053	-0.001	-0.035	-0.007	-0.022	-0.033
D27	-0.071	-0.077	-0.035	-0.04	-0.074	-0.047	-0.044	-0.047	-0.032	0.189	-0.006	-0.047	-0.014	-0.047	-0.054
D28	-0.076	-0.083	-0.041	-0.031	-0.074	-0.052	-0.05	-0.051	-0.039	.277(*)	0.017	-0.033	-0.026	-0.045	-0.061
D29	0.092	0.102	-0.048	0.192	0.146	-0.057	-0.047	-0.055	-0.029	-0.004	0.246	.302(*)	-0.027	0.024	-0.032
D30	-0.034	-0.035	-0.017	-0.027	-0.036	-0.023	-0.021	-0.019	-0.009	-0.012	-0.024	-0.034	-0.019	-0.011	-0.023
D31	-0.052	-0.057	-0.033	-0.047	-0.053	-0.035	-0.023	-0.034	-0.033	-0.062	-0.043	-0.052	-0.033	-0.045	-0.017
D32	-0.048	-0.052	-0.043	0.045	-0.019	-0.052	-0.04	-0.049	-0.043	-0.055	0.018	0	-0.044	-0.027	-0.018
D33	-0.061	-0.068	-0.055	-0.054	-0.062	-0.04	-0.033	-0.038	-0.037	-0.066	-0.049	-0.067	-0.036	-0.055	-0.026
D34	-0.103	-0.112	-0.058	-0.076	-0.109	-0.065	-0.044	-0.064	-0.061	-0.109	-0.074	-0.097	-0.05	-0.079	-0.022
D35	-0.072	-0.08	-0.043	-0.05	-0.077	-0.047	-0.037	-0.047	-0.045	-0.077	-0.05	-0.071	-0.044	-0.059	-0.009
D36	-0.13	-0.144	-0.076	-0.108	-0.137	-0.085	-0.059	-0.084	-0.079	-0.144	-0.038	-0.118	-0.08	-0.097	-0.032
D37	-0.063	-0.07	-0.034	-0.06	-0.071	-0.039	-0.029	-0.038	-0.036	-0.058	-0.048	-0.064	-0.036	-0.051	-0.016
D38	-0.029	-0.032	-0.018	0.024	-0.032	-0.045	-0.04	-0.044	-0.04	-0.059	0.037	0	-0.025	-0.04	-0.042
D39	-0.064	-0.072	-0.039	-0.048	-0.064	-0.043	-0.034	-0.043	-0.042	-0.071	-0.048	-0.067	-0.04	-0.057	-0.022
D40	-0.08	-0.089	-0.042	-0.053	-0.084	-0.052	-0.04	-0.05	-0.045	-0.08	-0.056	-0.075	-0.048	-0.052	-0.032
D41	-0.131	-0.146	-0.076	-0.11	-0.14	-0.082	-0.06	-0.084	-0.079	-0.146	-0.095	-0.12	-0.081	-0.098	-0.036
D42	-0.116	-0.129	-0.074	-0.097	-0.125	-0.081	-0.054	-0.082	-0.074	-0.137	-0.079	-0.1	-0.079	-0.08	-0.022
D43	-0.062	-0.069	-0.037	-0.05	-0.067	-0.041	-0.029	-0.04	-0.036	-0.062	-0.04	-0.056	-0.039	-0.034	-0.011
D44	-0.061	-0.065	-0.038	-0.049	-0.067	-0.042	-0.032	-0.042	-0.04	-0.071	-0.038	-0.054	-0.042	-0.044	-0.015
D45	-0.026	-0.024	-0.018	-0.002	-0.02	-0.021	-0.008	-0.022	-0.019	-0.043	-0.006	-0.007	-0.023	0.01	0.032
D46	-0.141	-0.156	-0.081	-0.115	-0.147	-0.092	-0.068	-0.09	-0.086	-0.154	-0.108	-0.138	-0.087	-0.108	-0.041

D47	-0.046	-0.051	-0.025	-0.046	-0.053	-0.028	-0.023	-0.027	-0.027	-0.047	-0.034	-0.049	-0.025	-0.037	-0.013
D48	-0.047	-0.051	-0.028	0.001	-0.043	-0.03	-0.026	-0.034	-0.011	0.161	0.033	-0.001	0.014	-0.023	-0.038
D49	0.051	0.071	-0.034	0.216	0.12	-0.041	-0.033	-0.039	-0.023	0.026	0.209	0.228	-0.015	0.037	-0.017
D50	-0.033	-0.035	-0.04	0.075	0.002	-0.05	-0.042	-0.048	-0.042	-0.04	0.049	0.028	-0.04	-0.018	-0.029
D51	-0.003	0.001	-0.056	0.215	0.052	-0.066	-0.057	-0.063	-0.045	0.014	0.24	0.18	-0.038	0.009	-0.047
D52	0.078	0.077	-0.046	0.095	0.119	-0.052	-0.041	-0.051	-0.034	0.043	0.188	0.189	-0.029	0.026	-0.043
D53	-0.037	-0.039	-0.024	0.03	-0.025	-0.03	-0.029	-0.03	-0.016	0.166	0.091	-0.001	-0.001	-0.021	-0.036

	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30
D16	1														
D17	-0.033	1													
D18	-0.024	-0.02	1												
D19	-0.001	0.002	0.242	1											
D20	-0.034	-0.027	-0.028	-0.021	1										
D21	-0.002	-0.052	-0.021	0.013	-0.049	1									
D22	0.002	-0.029	-0.028	-0.018	-0.027	-0.017	1								
D23	-0.012	-0.053	-0.015	0.061	-0.058	0.154	-0.017	1							
D24	-0.029	-0.028	-0.025	-0.019	-0.025	-0.044	0.029	-0.052	1						
D25	-0.032	-0.003	0.123	.450(**)	-0.04	-0.04	-0.039	-0.02	-0.038	1					
D26	0.01	-0.03	-0.025	-0.017	-0.027	-0.021	0.057	-0.028	-0.008	-0.04	1				
D27	-0.046	-0.051	-0.047	-0.033	-0.044	-0.056	0.068	-0.062	-0.037	-0.067	-0.02	1			
D28	-0.047	-0.057	-0.05	-0.041	-0.049	-0.057	0.16	-0.064	-0.035	-0.076	-0.009	.703(**)	1		
D29	-0.008	-0.051	-0.01	0.043	-0.057	0.089	-0.032	0.173	-0.052	-0.007	-0.037	-0.007	-0.004	1	
D30	-0.024	-0.025	-0.024	-0.02	-0.022	-0.028	-0.017	-0.034	-0.019	-0.036	-0.016	-0.013	-0.006	-0.038	1
D31	-0.04	0.179	-0.003	0.057	-0.032	-0.061	-0.038	-0.052	-0.036	0.045	-0.039	-0.065	-0.073	-0.045	-0.032
D32	-0.038	-0.032	-0.014	0.053	-0.046	-0.038	-0.045	-0.027	-0.05	0.009	-0.05	-0.075	-0.072	0.056	-0.04
D33	-0.043	-0.023	-0.017	0.021	-0.032	-0.064	-0.041	-0.071	-0.037	-0.007	-0.041	-0.068	-0.076	-0.054	-0.034
D34	-0.073	-0.033	-0.005	0.118	-0.052	-0.107	-0.071	-0.111	-0.065	0.033	-0.069	-0.117	-0.13	-0.086	-0.059
D35	-0.053	-0.028	-0.005	0.071	-0.041	-0.076	-0.051	-0.071	-0.047	0.019	-0.05	-0.084	-0.094	-0.058	-0.041
D36	-0.096	-0.033	0.009	0.209	-0.06	-0.139	-0.094	-0.138	-0.086	0.061	-0.092	-0.155	-0.174	-0.107	-0.079
D37	-0.046	-0.025	-0.005	0.062	-0.033	-0.066	-0.042	-0.071	-0.038	0.019	-0.041	-0.069	-0.078	-0.058	-0.034
D38	-0.038	-0.031	-0.033	-0.01	-0.041	-0.037	-0.033	-0.026	-0.04	-0.04	-0.034	-0.048	-0.046	0.076	-0.031
D39	-0.047	-0.024	-0.013	0.041	-0.036	-0.069	-0.043	-0.076	-0.042	0.002	-0.044	-0.075	-0.084	-0.053	-0.038
D40	-0.06	-0.028	-0.012	0.071	.360(**)	-0.086	-0.054	-0.091	-0.051	0	-0.055	-0.088	-0.099	-0.072	-0.045
D41	-0.098	-0.037	0.008	0.199	-0.05	-0.141	-0.094	-0.143	-0.086	0.066	-0.093	-0.156	-0.174	-0.114	-0.079
D42	-0.093	-0.025	0.035	0.251	-0.05	-0.131	-0.092	-0.127	-0.085	0.095	-0.091	-0.152	-0.17	-0.094	-0.076
D43	-0.046	-0.021	-0.011	0.08	-0.032	-0.066	-0.043	-0.068	-0.041	0.026	-0.044	-0.074	-0.082	-0.055	-0.036
D44	-0.049	-0.024	-0.001	0.079	-0.041	-0.071	-0.046	-0.066	-0.045	0.024	-0.046	-0.08	-0.09	-0.049	-0.041
D45	-0.024	-0.011	0.024	0.082	-0.03	-0.035	-0.032	-0.016	-0.031	0.082	-0.031	-0.054	-0.063	0.001	-0.027
D46	-0.103	-0.037	0	0.172	-0.061	-0.15	-0.1	-0.152	-0.091	0.043	-0.098	-0.165	-0.184	-0.12	-0.084
D47	-0.032	-0.024	-0.01	0.015	-0.022	-0.048	-0.028	-0.053	-0.026	0.001	-0.028	-0.047	-0.053	-0.048	-0.023
D48	-0.025	-0.041	-0.03	-0.009	-0.035	-0.014	0.053	-0.023	-0.03	-0.046	-0.018	0.17	0.202	-0.008	-0.008
D49	-0.003	-0.037	-0.002	0.031	-0.042	0.084	-0.013	0.151	-0.038	-0.003	-0.026	0.015	0.026	.773(**)	-0.02
D50	-0.035	-0.038	-0.027	0.011	-0.045	-0.02	-0.036	-0.005	-0.046	-0.019	-0.044	-0.06	-0.053	0.108	-0.033
D51	-0.035	-0.061	-0.03	0.021	-0.064	0.041	-0.014	0.077	-0.058	-0.039	-0.042	-0.026	-0.001	.407(**)	-0.011
D52	0.001	-0.048	-0.015	0.034	-0.053	0.097	0.022	0.097	-0.035	-0.025	0	-0.052	-0.044	0.14	-0.026
D53	-0.018	-0.036	-0.027	-0.02	-0.032	-0.008	0.095	-0.004	-0.02	-0.046	0.003	0.183	.288(*)	0.031	0.006

	D31	D32	D33	D34	D35	D36	D37	D38	D39	D40	D41	D42	D43	D44	D45
D31	1														
D32	0.05	1													
D33	0.048	0.096	1												
D34	0.132	0.257	0.227	1											
D35	0.081	0.162	0.139	.312(*)	1										
D36	0.218	0.246	0.247	.426(**)	.336(*)	1									
D37	0.057	0.065	0.06	0.162	0.117	.373(**)	1								
D38	-0.029	0.102	-0.03	-0.043	-0.039	-0.059	-0.04	1							
D39	0.065	0.132	0.12	.273(*)	0.194	.276(*)	0.081	-0.036	1						
D40	0.075	0.075	0.075	0.161	0.104	0.269	0.081	-0.042	0.088	1					
D41	0.207	0.253	0.263	.443(**)	.332(*)	.629(**)	0.266	-0.064	.281(*)	.341(*)	1				
D42	0.255	.281(*)	0.218	.444(**)	.325(*)	.582(**)	0.254	-0.046	0.263	.352(**)	.654(**)	1			
D43	0.078	0.089	0.073	0.183	0.119	0.267	0.08	-0.03	0.097	0.14	.310(*)	.363(**)	1		
D44	0.038	0.039	0.033	0.097	0.057	0.168	0.068	-0.022	0.048	0.038	0.132	0.143	0.048	1	
D45	0.066	0.068	0.039	0.147	0.084	0.254	0.079	0.007	0.063	0.061	0.209	0.247	0.075	.358(**)	1
D46	0.229	.336(*)	.345(*)	.565(**)	.427(**)	.666(**)	.280(*)	-0.061	.370(**)	.309(*)	.681(**)	.654(**)	.303(*)	0.148	0.226
D47	-0.016	-0.02	-0.019	-0.022	-0.021	-0.012	-0.014	-0.032	-0.01	-0.018	-0.023	-0.02	-0.013	-0.016	-0.008
D48	-0.051	-0.048	-0.054	-0.089	-0.059	-0.119	-0.054	-0.025	-0.057	-0.069	-0.119	-0.115	-0.056	-0.061	-0.039
D49	-0.037	0.084	-0.043	-0.062	-0.043	-0.074	-0.043	0.114	-0.04	-0.046	-0.083	-0.064	-0.037	-0.033	-0.001
D50	0.002	.914(**)	0.033	0.138	0.077	0.132	0.015	0.132	0.058	0.027	0.133	0.157	0.033	0.011	0.021
D51	-0.051	.273(*)	-0.048	-0.058	-0.043	-0.08	-0.055	0.185	-0.044	-0.059	-0.087	-0.056	-0.047	-0.04	0.009
D52	-0.045	-0.024	-0.054	-0.092	-0.07	-0.118	-0.061	-0.023	-0.056	-0.079	-0.122	-0.111	-0.055	-0.06	-0.023
D53	-0.045	-0.022	-0.047	-0.079	-0.056	-0.105	-0.048	0.032	-0.052	-0.061	-0.107	-0.099	-0.049	-0.05	-0.03

	D46	D47	D48	D49	D50	D51	D52	D53
D46	1							
D47	-0.025	1						
D48	-0.127	-0.036	1					
D49	-0.087	-0.034	0.043	1				
D50	0.193	-0.021	-0.032	0.144	1			
D51	-0.078	-0.053	0.042	.546(**)	.349(*)	1		
D52	-0.13	-0.046	-0.038	0.114	-0.006	0.059	1	
D53	-0.111	-0.033	.286(*)	0.117	0.003	0.176	-0.016	1