한국심리학회지 : 실험 및 인지
The Korean Journal of Experimental and Cognitive Psychology
2002, Vol. 14, No. 4, 309 - 319

뇌와 인지 모형: 잠재의미분석을 사용한 문서분류*

김 청 택† 이 태 헌

서울대학교 심리학과

본 연구의 목적은 인간이 산출한 문서들의 자료를 분석하여 그 의미를 파악하고 파악된 의미를 이용하여 문서를 분류하는 기법을 개발하는 것이었다. 한 단어의 의미를 그 단어와 동시에 발생하는 문서들의 패턴으로 정의하고, 단어와 문서들의 수반성 표를 분석함으로써 의미공간을 수학적으로 표상하였다. 이 공간상에서 문서를 표상하여 문서들간의 상관들을 분석함으로써 문서분류시스템을 개발하였다. 이때 사용된 방법은 Singular Value Decomposition 방법에서 축회전을 도입한 방법이었다. 이에 의한 문서분류의 결과는 인간의 분류와 유사한 수행을 보였으며, 축회전기법을 도입하였을 때 인간의 문서분류와 보다 근접하는 결과를 보였다.

주제어 잠재의미분석, 문서분류, 맥락, 신경망

* 이 연구는 과학기술부의 뇌과학 연구사업의 지원을 받았음.

† 교신저자 : 김 청 택, (151-746) 서울시 관악구 신림동, 서울대학교 심리학과,
E-mail : ctkim@snua.c.kr)

언어처리에 관한 인지심리학 연구결과들에 따르면 단어가 개별적으로 존재할 때는 그 의미가 확정되지 않으며, 단어의 의미는 맥락(context)이 주어질 때 비로소 파악된다(Forster, 1979; Sweeney, 1979). 언어처리에서 맥락의 효과는 거듭하여 강조되어왔고 언어처리과정을 설명하는 주요한 설명개념으로 사용되어 왔다. 그러나 맥락자체에 대한 연구는 드물며 여러 연구들에서 사용되는 맥락이라는 용어는 연구자에 따라 각기 다르게 사용되었다. 이 연구에서는 맥락에 대한 조작적 정의를 내리고 이 정의에 따라서 맥락이 단어나 문장, 텍스트의 의미 결정에 어떠한 영향을 미치는지를 연구하고자 하였다.

본 연구에서 맥락은 파악하고자 하는 단어/문장/텍스트 등과 동시에 발생하는 즉 공기하는(co-occur) 자극의 패턴으로 정의하였다. 이러한 정의는 Landauer와 Dumais(1996, 1997)의 잠재 의미분석(Latent Semantic Analysis, 이하 LSA)에서 단어의 의미가 그 단어와 동일한 문장에서 제시된 다른 단어들과의 공기성 정보, 즉 맥락에 의해 결정된다고 가정하는 것과 동일하다. 이러한 공기성에 의한 맥락의 정의는 연합주의 전통에 따르면 그리 무리한 정의는 아니며, 신 연합주의에서 동시 발생적인 자극들의 연결을 강화시키고 그 연결 패턴에 의하여 의미가 파악되는 것은 Hebb의 학습규칙에 기반을 둔 신경망의 원리에서도 이미 적용되고 있는 것이다. Myung, Kim과 Levy (1997)는 Hebb의 학습원리(Hebb, 1949)를 기반으로 하는 신경망을 구성하여 자극들의 공기성 정보가 지각과정에서 맥락의 효과를 모사할 수 있음을 보여주었다. 또한 잠재의미분석은 단어의 의미가 이러한 공기성 정보를 분석함에 의하여 결정될 수 있음을 보여주었다.

이 논문에서는 이러한 LSA를 확장하여 한 단어의 의미를 그 단어가 제시된 문서들의 집합으로 정의하고, 한 문서의 의미를 그 문서 속에 나타난 단어들의 집합으로 정의하였다. 그리고 문서들의 의미를 분석함에 의하여 문서들간의 유사성을 계산하고 이에 의하여 문서들을 범주화 혹은 분류할 수 있는 시스템 즉 인공의 뇌 시스템을 구축하는 방법을 제시하였다. 이 논문은 구성은 다음과 같다. 먼저 잠재의미분석에 대한 개관을 하고, 이 잠재의미분석을 확장하여 새롭게 해석하여 분석분류에 적용시킬 수 있다는 것을 보였으며, 마지막으로 시뮬레이션 연구를 통하여 이 방법의 타당성을 보여주었다.

잠재의미분석

잠재의미분석은 단어와 맥락의 수반성을 분석함으로써 의미구조를 파악하는 방법이다. 먼저 한 단어의 의미는 그 단어가 각각의 맥락(예컨대 문장) 속에서 제시된 빈도에 의해 표상된다. 예컨대 50개의 단어들이 3가지 맥락 속에서 제시되는 경우를 생각하여 보자. 즉 50개의 단어들은 세 개의 문장 속에서만 제시되는 경우이고 우리에게 제시된 문장은 세 개 밖에 없는 경우이다. 이때 각 단어들은 3차원으로 구성된 맥락 속에서 제시된 빈도로서 표시된다. A라는 단어가 세 맥락(문장)에서 각각 2번, 3번, 1번씩 출현하였다면 이 단어는 3차원 공간상에서(2,3,1)의 한 벡터 혹은 점으로 표시된다. 이렇게 표시된 공간의 예가 그림 1에 제시되어 있다.

실재로 우리가 접하는 환경에서는 그림 1의 공간은 수만 차원에서 수백만 차원까지 확장되어야 할 것이다. 이를 의미공간이라 하자. 이 의미공간에서 두 단어간의 의미 유사성은 두 단어의 의미 벡터의 cosine값(혹은 상관계수)으로 정의될 수 있다. 잠재의미분석은 이러한 의미공간을 그대로

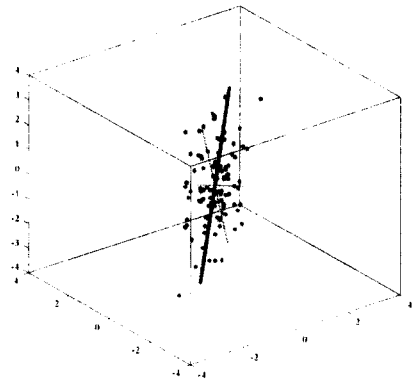

그림 1. 다차원 공간의 의미표상과 주축들
50개의 단어들이 3가지 맥락 속에서 제시되는 예

사용하는 것이 아니라 차원을 축소하여 새로운 의미관계를 산출하여 내는 방법이다. 이러한 차원 축소 방법은 요인분석이나 MDS, 주성분분석 등의 기본을 이루는 선형 대수학의 정리인 singular value decomposition 정리(이하 SVD)를 사용하였다. 보다 구체적으로 원래의 수반성 자료를 X라 하자. 원자료 행렬은 SVD의 정리에 의하여 항상

$$X = UDV'$$

로 재기술될 수 있다. 이때 U와 V의 각 행은 서로 직교행렬로서 각각 맥락과 단어의 축을 나타내며, D는 대각행렬로 각 요소는 설명분산을 나타내게 된다. 이때의 축은 원래의 축이 아니고 설명분산에 따라 재배열된 축이다. 그림 1에서 새로운 세 가지 축이 제시되어 있는데, 가장 굵은 선의 첫 번째 축이 단어들 분산/공분산의 가장 많은 부분을 설명하고, 두 번째 축, 세 번째 축으로 감에 따라 설명 분산이 줄어드는 것을 알 수 있다. LSA에서는 차원의 축소에 의한 맥락의 추상화를 유도하는 데, 그림 1의 예에서 첫 번째 축 만을 사용하여 단어들은 표시하면 3차원에서 1차원으로 차원이 감소하게 된다. 이렇게 축소된 축들을 사용하여 다시 의미공간을 구성하면, 새로운 수반성 행렬 $\hat{X}$는

$$\hat{X} = U_r D_r V_r'$$

이 된다. 여기에서 U_r, V_r, D_r은 각각 축소된 공간상에서의 맥락과 단어 축과 각 축의 설명분산을 나타내게 된다. Landauer와 Dumais(1997)는 감소된 차원을 사용하여 단어의 의미를 표상하였다. 여기에서 흥미로운 점은 원래의 축을 모두 사용하는 경우보다, 축소된 공간을 사용하는 경우에 인간의 의미판단과 더 유사한 결과를 얻었다는 점이다. 일반적으로 차원의 축소는 축소 전에 있던 정보의 손실을 유발하게 되므로 공간의 축소가 도리어 수행을 향상시킨다는 점은 쉽게 이해되지 않는다. 그러나 차원의 축소는 단어의 의미를 구성하고 있는 맥락의 요약 내지는 추상화를 유도함으로써 인간의 인지과정과 유사한 측면을 띠게 되고 이러한 점이 인간의 수행과 더 유사한 결과를 내었을 것이다.

Landauer와 Dumais(1997)는 LSA에 의한 의미표상을 이용하여 유사어 찾기와 같은 의미처리를 할 수 있음을 보여주었다. 그들은 백과사전에 나와 있는 총 30,473개의 문서와 적어도 두 개 이상의 문서에서 출현한 60,768단어를 이용하여 300차원상에서 LSA를 실시한 결과를 이용하여 TOEFL에서 사용되었던 동의어 검사를 실시하였다. 이때 LSA에서 동의어를 선택하는 방법은 문제에 제시된 단어와 보기에 제시된 단어의 축소된 차원상에서의 상관값을 구하여 가장 높은 상관을 가지는 단어를 정답으로 선택하였다. 그 결과 총 80문제 중 64.4%의 정답률을 보였으며 이는 영어를 모국어로 하지 않는 외국인이 미국 대학에 지원하기 위해 실시한 TOEFL 시험 자료에 축적된 동

의어 검사의 평균 정답률인 64.5%에 필적할 만한 결과였다.

그러나 이러한 시뮬레이션 결과는 매우 수학적이고 기계적인 처리과정을 통해 나온 것이기 때문에 그 의미를 해석하기가 용이하지는 않다. 즉 차원을 축소하는 과정이 실제 인간의 언어처리과정에서 일어나는지를 알 수 없으며, 단어 혹은 맥락이 지니고 있는 숫자들의 집합 즉 벡터가 무엇을 가리키는 것인지는 정확히 알 수 없다. 그러나 무시할 수 없는 점은 LSA가 인간의 언어 습득 혹은 이해 기능을 모사하는 결과를 산출해 낼 수 있다는 점이다.

잠재의미 분석을 통한 문서분류

본 연구의 주목적은 이러한 LSA를 확장하여 문서를 분류할 수 있는 시스템을 설계하는 것이었다. LSA는 단어 즉, 행벡터를 기준으로 단어의 의미를 표상한다. 그리고 문장의 의미는 문장 속에 포함되어 있는 단어들의 벡터 합으로서 정의하였고, 문서의 의미는 문장들의 벡터 합으로서 정의되었다. LSA의 맥락 속에서 문서를 분류하는 방법은 문장들의 벡터 합으로 구성된 문장들의 상관행렬을 분석함으로써 이루어 질 수 있을 것이다. 그러나 이 방법은 벡터의 합을 계산하는 것은 많은 시간이 요구된다는 점에서 효율적이지 못하다.

LSA가 간과하고 있는 점은 SVD 기법은 수학적으로 행과 열을 동시에 분석할 수 있는 기법이라는 점이다. 즉 단어의 차원을 축소하고 축을 변화시키는 것은 그와 동시에 맥락의 차원를 축소하고 축을 변화시키는 것을 수학적으로 함의하기 때문이다. 다시 말하면 차원의 축소와 변화가 열벡터에게도 이미 적용되어 있다는 것이다(LSA에서 열벡터의 분석은 맥락 즉 문서에 대한 분석이며 행벡터의 분석은 단어에 대한 분석이다). 즉 SVD에서 V'행렬의 행이 X행렬의 행에 대한 새로운 직교 축을 형성하므로 SVD는 행벡터에 대한 구조뿐만 아니라 열벡터에 대한 분석을 동시에 할 수 있다. 위의 예에서 문서를 50차원 상에서 3개의 점(points)으로 표현한 것으로 기하학적 해석을 가한다면, 문서간 관계성 역시 파악할 수 있으며, SVD을 하는 과정에서 50개의 차원이 이미 축소되어져 있고 축이 변형되어 있다는 것이다. 단어를 축으로 하고 문서가 점으로 표시된 공간에서 벡터들의 군집이 특정 패턴을 이루고 있다면 그 패턴을 잘 설명 할 수 있는 몇 개의 차원만을 선택함으로써 여러 문서들이 이루고 있는 고차적 상관관계를 좀 더 쉽게 파악할 수 있다.

그런데 이러한 SVD 기법을 문서 분류에 응용하기 위해서는 단순히 행렬을 SVD으로 분해한 후 차원을 축소하고 다시 재결합한 결과만으로는 부족하다. 왜냐하면, SVD에서 사용되는 차원 축소의 기준은 자료의 분산을 가장 많이 설명해 주는 차원을 선택하는 것에서 그칠 뿐 선택된 차원의 의미가 어떤 것인가에 대한 것은 알 수 없다. 즉 공학적으로 인간의 행동을 흉내내는 데에는 도움이 되겠지만, 각 축이 의미하는 바를 추출하는 것은 쉽지 않다. 또한 문서간 유사성을 확인하기 위해 축소된 차원에서 문서들이 가리키고 있는 벡터들의 단순 상관을 모두 구한다는 것도 계산의 부하량에서 무리가 따르면, 단순 상관들을 모두 구한다 하더라도 고차적 관계성은 결코 파악되지 않는다. 이러한 이유로 문서 분류 시뮬레이션에서는 SVD에서 축소된 수반성 행렬($\hat{X}$)를 그대로 사용하지 않고, 새롭게 생성된 차원과 문서 벡터들간의 상관값을 계산하여 사용하였다. 이 상관값은 D_r행렬과 V_r'행렬의 곱으로 계산

될 수 있다. 즉

$$Cov(U_r, X) = E(U_r' U_r D_r V_r' + U_r \delta) = D_r V_r'$$

$$where \quad X = U_r D_r V_r' + \delta$$

이 두 행렬의 곱으로 구성된 행렬의 행벡터들은 각 차원이 여러 문서들과 가지는 상관값을 나타낸다. 따라서 상관값이 높은 문서들을 하나의 군집으로 묶을 수가 있으며, 이렇게 묶인 문서들은 공통된 내용을 지닌 것으로 간주할 수 있을 것이다.

연구 1 : 문서분류 시뮬레이션

연구 1은 실제 자료를 사용하여 위에서 제시된 방법의 타당성을 검증하기 위하여 행하여 졌다. 서울대학교 심리학과 게시판에 게시된 웹문서를 사용하여 SVD에 기반한 문서분류를 행하고 이 분류가 합당한지를 관찰하였다.

방법

문서 자료는 2000년 8월 24일부터 2002년 7월 16일까지 서울대학교 심리학과 게시판에 등록된 1748개의 문서들이었다. 원자료 X 행렬은 적어도 두 문서 이상에서 출현한 단어 19,029단어를 대상으로 문서별 출현 빈도를 계산하여 만들어졌다. 엔트로피 측정치를 이용하여 각 단어의 정보가를 계산한 새로운 행렬을 기본 구조로 분해한 뒤 총 9개의 차원을 선택하여[1] $D \times V'$ 행렬을 계산하였다. 관례상 $D \times V'$의 전치행렬을 사용하므로 여기서도 전치된 행렬을 제시하였다. 표 1은 $D \times V'$의 전치 행렬의 일부분을 발췌한 것으로 각 차원에서 상관값이 가장 높았던 문서들 20개씩을 추출한 것이다. 이 표에서 상관값이 높은 문서들은 하나의 군집으로 묶을 수 있으며 이는 곧 동질적인 내용을 지닌 문서들이라는 가정을 확인해 준다.

결과 및 논의

표 1에는 9개의 차원 각각과 문서들이 이루는 상관값이 제시되어 있다. 맨 윗줄의 1부터 9 가지의 숫자는 9개의 새로운 축을 의미하며 각 축 내에 있는 두 개의 열들은 각각 문서 번호와 상관값을 의미한다. 굵은 글씨로 표시되어 있는 숫자들은 상관값이 상대적으로 높은 문서들을 가리키고 있는데, 이 문서들의 내용을 분석한 결과 매우 유사한 문서들이었다. 상관값들이 매우 높은 문서들은 인터넷 게시판의 특성을 잘 반영하듯 동일한 광고 문서들이었다. 또한 각 차원에서 상관값이 높은 문서들의 내용을 분석해 보면, 각 차원들은 심리학과 편입 및 대학원 입학문의, 광고 및 홍보, 졸업후 진로문의, 대학원 스터디 모집광고, 기관 및 단체 워크샵 홍보, 과외 사이트 홍보, 심리학 관련 유학 문의 등의 내용을 담고 있었다.

이러한 결과를 종합해 보면, LSA 모델은 단어 벡터들의 군집 패턴을 분석할 수 있을 뿐만 아니라, 문서 벡터들의 군집 패턴 역시 축소된 차원에서 효과적으로 분석 할 수 있으며, 인간의 문서 분류와 유사하게 문서의 내용을 중심으로 분류를 해 낸다는 것을 알 수 있다. 그러나 문서

1) 9개의 차원을 선택할 때는 scree test를 이용하였다. 이는 singular value를 차원을 x축으로 하여 그래프로 나타낸 뒤, 곡선의 기울기 감소가 급격히 떨어지는 지점에서 차원의 수를 결정하는 방법이다.

표 1. 문서분류의 결과 (각 축에서 첫 번째 열은 문서번호를 두 번째 열은 축과 문서의 상관을 나타낸다).

1		2		3		4		5		6		7		8		9	
174	-0.17	72	0.26	183	-0.05	325	0.06	157	0.02	364	-0.29	41	-0.35	72	-0.14	569	-0.24
286	-0.20	278	0.23	782	-0.04	722	0.16	278	0.74	365	-0.29	375	-0.35	521	-0.15	578	-0.20
394	-0.19	323	0.20	783	-0.04	1225	0.06	323	0.63	451	-0.57	878	-0.30	951	-0.11	581	-0.21
538	-0.17	336	0.25	934	-0.03	1405	0.22	336	0.83	453	-0.29	1076	-0.35	955	-0.11	646	-0.25
566	-0.18	424	0.21	939	-0.04	1462	0.08	566	0.02	462	-0.59	1132	-0.36	1006	-0.15	941	-0.20
581	-0.18	451	0.28	941	-0.04	1487	0.20	622	0.83	501	-0.59	1453	-0.57	1310	-0.18	943	-0.19
646	-0.21	462	0.30	943	-0.04	1492	0.07	743	0.82	518	-0.59	1539	-0.59	1311	-0.19	1403	-0.41
742	-0.19	501	0.29	984	-0.04	1566	0.87	772	0.83	1453	-0.22	1576	-0.29	1312	-0.18	1406	-0.42
798	-0.20	518	0.29	985	-0.03	1571	0.87	937	0.03	1539	-0.23	1586	-0.40	1313	-0.16	1444	-0.41
861	-0.18	521	0.26	1310	-0.03	1572	0.84	1348	0.02	1643	-0.23	1643	-0.59	1403	-0.77	1445	-0.39
1094	-0.17	622	0.25	1348	-0.03	1604	0.85	1367	0.03	72	0.31	364	0.28	1406	-0.80	72	0.31
1338	-0.17	743	0.27	1368	-0.03	1610	0.12	1377	0.03	244	0.19	365	0.28	1444	-0.77	505	0.13
1340	-0.18	772	0.24	1445	-0.03	1614	0.64	1483	0.03	521	0.32	451	0.46	1445	-0.71	508	0.13
1483	-0.21	1006	0.26	1483	-0.03	1621	0.21	1621	0.10	855	0.18	453	0.28	1453	-0.12	521	0.34
1566	-0.29	1476	0.26	1621	-0.84	1630	0.79	1635	0.20	1006	0.33	460	0.15	1465	-0.13	950	0.19
1571	-0.29	1621	0.47	1637	-0.84	1637	0.21	1637	0.10	1213	0.18	462	0.48	1466	-0.13	951	0.23
1572	-0.29	1637	0.47	1645	-0.84	1645	0.21	1645	0.10	1476	0.33	487	0.16	1475	-0.14	955	0.23
1604	-0.31	1645	0.47	1649	-0.84	1649	0.21	1649	0.10	1500	0.19	501	0.47	1476	-0.15	963	0.19
1614	-0.26	1649	0.47	1698	-0.84	1666	0.06	1692	0.09	1518	0.20	518	0.47	1539	-0.12	1006	0.34
1630	-0.29	1698	0.47	1716	-0.09	1698	0.21	1698	0.10	1548	0.20	712	0.14	1643	-0.12	1476	0.34

분류 결과를 보다 세밀하게 분석해 보면, 내용의 유사성을 중심으로 문서들이 묶여지긴 했지만, 이 결과를 해석하기는 그다지 용이하지 않았다. 특히 두 번째 차원과 세 번째 차원은 동일한 문서들과의 상관값이 매우 높아 굳이 차원을 구분할 이유가 없는 듯 보였다. 또한 6,7,9 번째 차원은 반대 방향을 가리키는 문서 벡터들이 같은 차원에서 높은 상관값을 가지고 존재함으로써 해석이 더욱 어려워졌다. 이렇게 같은 차원에서 반대 방향을 가리키고 있는 문서들은 그 내용이 대조적인 것인지 아니면 전혀 무관한 것인지는 파악하기가 매우 어렵다. 예를 들어 7번째 차원에서 음의 상관값을 가지는 문서들은 과외 사이트를 홍보하는 글이었음에 반해, 양의 상관값을 가지는 문서들은 OO 대학 특별 전형 홍보와 OO대학 토요 세미나 일정 홍보에 관한 내용이었는데, 이 내용들은 대조적인 것이라기 보다는 오히려 홍보라는 동질적 범주로 묶일 수도 있었을 것이며 또는 전혀 다른 차원에서 각각이 묶여 있었다면 오히려 해석하기가 더 쉬워졌을 것이다.

연구 2 : SVD의 불확정성: 축의 회전

앞서 제시되었던 문서 분류 결과의 문제점에 대한 해결책은 LSA의 출발점에서부터 찾을 수 있

다. 즉 애초의 축을 변형하여 새로운 축을 생성하고 자료의 분산을 가장 잘 설명해 주는 9개의 축을 선택했듯이, 이렇게 선택된 새로운 축들도 우리가 좀 더 해석하기 쉽도록 다시 변형할 수 있는 것이다. 동일한 문제점이 심리측정의 요인분석법에서도 나타나는데, SVD에 의한 요인 부하량 해의 비확정성이 그것이다. 요인분석법에서는 축의 회전시킴으로써 해석하기 좋은 구조를 찾아낸다 (Browne, 2001 참조). 이 연구에서는 이러한 축회전법을 사용하여 해석하기 위한 구조를 찾아내는데 목적을 두었다. 축의 회전은 U 행렬의 오른쪽에 적당한 회전행렬 T 행렬을 곱함으로써 이루어질 수 있다. T 행렬의 종류에 따라 축의 회전 이 결정되게 된다.

$$\hat{X} = U_r D_r V_r' = U^* V^{*\prime} = U^* T T' V^{*\prime} = U^+ V^{+\prime}$$

$$where\ U^* = U_r, \quad V^{*\prime} = D_r V_r',$$

$$U^+ = U^* T, \quad V^+ = V^* T, \quad T'T = I$$

위의 식에서 $\hat{X}$는 $U^* V^{*\prime}$으로 기술될 수 있는데 U^*와 V^*의 우측에 $T'T=I$가 되는 행렬을 구하여 곱하면 $\hat{X}$는 다시 동일한 구조를 지닌 $U^+ V^{+\prime}$으로 기술되어 있어서 동일한 $\hat{X}$를 산출하는 무한히 많은 U들과 V들이 존재하게 된다 (이는 직교회전에만 한정된다). 즉 축의 불확정성이 있게 된다. 이러한 불확정성을 이용하여 우리가 해석하기 좋은 형태로 U를 정하는 것이 회전에 해당되게 된다. 회전의 장점은 분석자가 원하는 데로 회전되는 방식을 정할 수 있으며, 이때 심리학적인 요소를 도입하면 인간에게 적절한 회전을 할 수 있게 된다. 이 연구에서는 이러한 회전법을 도입하여 연구 1에서 인간의 분류방법과 일치되지 않게 분류된 문서분류의 문제를 해결하였다.

방법

사용된 문서자료는 연구 1과 동일하였다. 회전행렬 T는 다음과 같은 방식으로 정하여졌다. 연구 1의 결과를 살펴보면(표 1) 각 차원에서 상관값이 낮은 문서들은 상대적으로 상관값들이 높은 문서 군집들과 내용이 다른 경우가 발견되었다. 이는 해석하기 힘든데, 해석하기 좋은 형태는 낮은 상관값들은 0에 가까워지고, 높은 상관값들은 양의 큰 값을 가지며, 동일 차원 내에서 높은 상관값을 가지지만 부호가 방향이 반대인 문서들은 다른 차원으로 분리되는 것이다. 이에 해당하는 회전 방식 중의 하나가 VARIMAX방식이므로 이 방식을 회전방식으로 채택하였다. 그리고 직교회전을 하였는데, 각각의 축들이 직교성을 유지하는 것은 다른 차원에서 묶여진 문서간에는 관련성이 없다는 것을 의미하므로 이 또한 분류된 내용을 해석하는데 도움이 될 것이다. 이러한 목적에 부합하는 결과를 산출하는 행렬 T를 구한 뒤 이 행렬을 이용하여 축을 회전한 뒤, 회전된 축과 문서 벡터들간의 상관값을 다시 계산하였다.

결과 및 논의

축이 회전된 다음 각 축과 문서들간의 상관값이 높은 20개의 문서들이 표 2에 제시되어 있다. 결과를 살펴보면 첫 번째 차원에 대부분의 심리학과 학사편입, 대학원입학, 유학 문의와 관련된 문서들이 대부분이 포함되었고, 두 번째 차원은 이런 질문에 대한 대답에 관한 문서들이 상대적으로 높은 상관값들을 가지면서 묶였다. 나머지 차원에서는 상품 광고, 대학원 스터디 모집, 과외 사이트 홍보, OO단체 워크샵 홍보, 유학관련 문의, OO대학 특별 전형 홍보 등의 문서들이 서로

표 2. 새롭게 찾은 축을 이용한 문서분류 결과(각 축에서 첫 번째 열은 문서번호를 두 번째 열은 축과 문서의 상관을 나타낸다)

1		2		3		4		5		6		7		8		9	
157	-0.18	226	0.17	128	-0.03	287	0.08	278	0.81	331	-0.08	41	-0.43	163	-0.09	11	0.17
286	-0.21	240	0.14	385	-0.03	325	0.08	323	0.69	364	-0.45	375	-0.43	203	-0.05	72	0.56
394	-0.18	257	0.15	420	-0.04	636	0.07	336	0.91	365	-0.46	506	-0.28	647	-0.05	147	0.18
538	-0.18	569	0.34	424	-0.10	722	0.19	448	0.03	420	-0.10	878	-0.34	656	-0.05	244	0.20
553	-0.18	576	0.19	559	-0.03	1225	0.09	548	0.03	451	-0.84	1016	-0.29	909	-0.05	281	0.17
566	-0.21	578	0.29	646	-0.03	1304	0.09	622	0.92	453	-0.46	1024	-0.27	1310	-0.20	349	0.17
651	-0.18	581	0.32	784	-0.03	1307	0.07	646	0.02	460	-0.30	1076	-0.42	1311	-0.20	521	0.59
742	-0.17	605	0.25	871	-0.03	1405	0.25	705	0.03	462	-0.87	1132	-0.42	1312	-0.20	855	0.20
798	-0.21	646	0.38	938	-0.03	1462	0.10	743	0.91	487	-0.30	1339	-0.32	1313	-0.18	951	0.19
951	-0.19	672	0.15	1096	-0.03	1487	0.25	772	0.90	501	-0.86	1380	-0.18	1403	-0.88	955	0.19
955	-0.21	934	0.15	1261	-0.06	1492	0.09	787	0.05	518	-0.86	1453	-0.66	1406	-0.91	963	0.17
963	-0.19	939	0.14	1576	-0.03	1566	0.95	853	0.03	569	-0.08	1539	-0.68	1444	-0.88	1006	0.60
1094	-0.19	941	0.16	1621	-1.00	1571	0.94	937	0.04	581	-0.08	1576	-0.38	1445	-0.82	1103	0.17
1335	-0.17	943	0.16	1626	-0.05	1572	0.91	1261	0.05	646	-0.10	1579	-0.36	1465	-0.07	1161	0.18
1338	-0.17	953	0.16	1637	-1.00	1604	0.92	1411	0.03	712	-0.27	1583	-0.19	1466	-0.07	1198	0.18
1348	-0.20	967	0.16	1645	-1.00	1610	0.15	1524	0.02	1588	-0.09	1586	-0.49	1475	-0.07	1213	0.20
1361	-0.17	984	0.16	1649	-1.00	1614	0.71	1635	0.24	1626	-0.13	1588	-0.30	1693	-0.07	1476	0.59
1367	-0.18	1020	0.15	1650	-0.03	1630	0.86	1692	0.12	1627	-0.09	1590	-0.36	1739	-0.12	1500	0.19
1428	-0.18	1103	0.15	1698	-1.00	1666	0.08	1714	0.03	1628	-0.09	1592	-0.30	1742	-0.06	1518	0.19
1483	-0.24	1594	0.15	1716	-0.13	1693	0.07	1737	0.03	1675	-0.15	1643	-0.68	1744	-0.05	1548	0.19

다른 차원에서 높은 상관값을 가지고 묶였다. 한 가지 특이할 만한 것은 7번째 차원은 HTML 태그를 사용한 문서들이 추출되었는데, 이는 단어-맥락 수반성 행렬을 만들 때, HTML 태그를 사용한 문서는 태그 역시 포함시켰기 때문에 이러한 문서들을 따로 분류한 결과로 해석될 수 있다.

축의 회전을 통해 얻어진 이러한 결과는 표 1의 결과에 비해 훨씬 더 해석하기가 용이하다. 우선 같은 차원 내에서 상관값이 높고 방향이 반대인 값들이 존재하지 않아 해석의 모호성이 줄어들었다. 또한 관련성이 높은 문서들이 동일한 범주에 분류되는 과정에서 각기 다른 내용은 서로 다른 차원에 배분됨으로써 각 차원의 의미를 해석할 수 있는 가능성을 열어 주고 있다. 또한 동일 차원 내에서 동질적인 문서들의 상관값들은 상대적으로 더욱 높아지고, 관련성이 적거나 전혀 다른 내용의 문서들의 상관값들은 0에 더욱 근접하게 됨으로써 역시 차원의 의미 해석에 도움을 주고 있다.

종합 논의

이 논문에서는 LSA을 확장하여 문서분류에 이

기법을 적용시키는 것에 목적으로 두었다. 여기에서 제안된 문서분류 모형은 문서가 단어들과 문장들의 벡터 합으로 표상되는 LSA와는 달리 SVD에서 문서들간의 유사성 정보를 지니고 있는 행렬을 직접 추출하고 이를 이용하여 문서를 분류하였다. 분류된 결과들은 인간이 하는 문서 분류와 유사하였다. 이는 인간의 의미표상 방식이 각 단어 혹은 문서의 맥락에 의하여 정의될 수 있음을 시사한다. 즉 한 단어의 의미가 Anderson (1990)에서 마디(node)로 표시되는 국소적인 표상을 가지고 있는 것이 아니라, 그 단어와 동시에 발생하는 맥락들에 의해 분산적으로 정의된다는 것이다. 이러한 분산표상은 대부분의 신경망에서 이미 가정되고 있는 것이며, 사실 LSA의 기법은 전환함수가 선형인 삼층의 피드포워트(feed-forward) 신경망과 수학적으로 유사하다. 많은 인공지능 모형에서 문장들의 분류라든지 단어의 의미 파악을 위하여 신경망을 사용하지만, 수만 혹은 수십만의 단위를 가진 신경망을 사용하기가 가능하지는 않은데, LSA의 기법은 비록 선형적이기는 하지만 아주 큰 신경망을 구성하는 것을 가능하게 하였다.

LSA의 심리학적 해석과 관련하여, 이 연구가 인지과정을 직접적으로 밝혀주지는 못한다는 주장을 할 수 있다. 실험심리학에서 가장 많이 사용하는 방법인 인지과정에 대한 가설을 마련하고 그 가설을 검증할 수 있는 조작을 인간피험자에게 가하는 방법론을 이 연구에서 사용하는 것은 아니다. 그러나 이 연구의 출발점은 자연스러운 상태에서 인간에 의하여 생성된 자료가 곧 인간의 인지과정을 밝히는 데 도움이 될 수 있다는 데 있다. 비록 실험적으로 조작하지는 않았더라도 사람들에 의해 자연스럽게 산출된 언어자료를 분석하고 이 분석에 의하여 인간의 인지과정을 밝히는 것도 또한 의미 있는 일일 것이다. 이러한 점에서 본 연구는 사람들이 생성한 자료에 의하여 추론된 의미구조와 이를 이용한 문서분류가 타당한지를 살펴보는 것이었다.

이 연구의 결과는 제안된 문서분류 모형이 인간의 사고 기능을 모사할 수 있다는 점에서도 중요하지만, 실질적으로 문서 검색 시스템에 활용될 수 있다는 점에서도 의의를 찾을 수 있다. 현재 많은 검색 시스템이 입력된 키워드를 지닌 문서를 모두 출력하는데 그치는 경우가 많은데 이러한 검색 결과는 비록 그 단어가 포함되어 있을지라도 전혀 의도하지 않은 문서들이 검색되는 단점이 있다. 그러나 이 모형에 의하여 문서를 분류한 뒤 검색된 키워드가 존재하는 문서 중 특정 차원과 가장 높은 상관값을 지니는 문서를 찾고 그 차원 내에서 동질적인 문서를 상관값이 높은 순서로 제공해 준다면 인간이 원하는 검색 결과를 얻을 수 있을 것이다.

그러나 LSA 모델을 이용하여 방대한 양의 문서들을 내용의 유사성을 기준으로 분류하여 검색 시스템에 응용할 수 있으려면, 좀 더 정교한 모델이 필요할 수도 있다. 왜냐하면 단순히 수학적 기제에 의해 새로운 차원을 찾아내고 또다시 수학적 기준에 의한 제약 조건 하에서 새로운 축을 찾아내는 것만으로는 그 결과에 대한 평가는 수학적으로만 가능할 뿐이지 분류 결과의 내용적 타당성은 판가름하기 힘들기 때문이다. 이는 LSA 모델을 통해 분류된 수 천, 수 만 개 문서들의 내용을 일일이 사람이 확인할 수 없기 때문에 더욱 그러하다.

이러한 측면에서 본다면, LSA 모델의 정교화는 수학적 제약만을 이용하기보다는 인간이 납득할 만한 기준, 혹은 결과물을 제공할 수 있는 방향으로 이루어져야 할 것이다. 이에 대한 한 가지 해결책으로서 생각해 볼 수 있는 것은 전혀 다른 제약 조건을 가진 새로운 T행렬을 이용하는 것

이다. 수학적으로 가능할 뿐만 아니라 문서 분류 결과의 타당성 혹은 인간이 납득할 만한 제약 조건으로서 다음과 같은 것을 생각해 볼 수 있을 것이다. 즉 의미상 묶여져야 할 문서들, 다시 말하면, 동일한 문서 혹은 동일한 내용의 문서들은 같은 축에서 매우 높은, 거의 1에 가까운 상관을 가지도록 하는 조건 하에서 T행렬을 생성할 수 있을 것이다. 또한 각 축들이 상호 직교할 필요가 없다는 조건 하에서 T행렬을 생성할 수도 있다. 나아서 위 두 조건을 동시에 만족할 수 있는 T행렬을 찾을 수 있다면, 좀더 그럴듯한 기준에 의해 문서를 분류하고 검색 시스템에 활용할 수 있을 가능성이 있다.

참고문헌

Anderson, J. R. (1983). The architecture of cognition. Cambridge, Mass: Harvard University Press.

Browne, M. W. (2001). An Overview of Analytic Rotation in Exploratory Factor Analysis. *Multivariate Behavioral Research*, 36, 111- 150.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In W. E. Cooper & E. Walker (Eds.). *Sentence processing: Psycholinguistic studies presented to Merrill Garret.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Landauer, T. K., & Dumais, S. T. (1996). How come you know so much? From practical problem to theory. In D. Hermann, C. McEvoy, M. Johnson, & P. Hertel (Eds.), *Basic and applied memory: Memory in context*. Mahwah, NJ: Erlbaum, 105-126.

Landauer, T. K. and Dumais S. T. (1997). A Solution to Plato's Problem: The Latent Semantic analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review, 104,* 211-240.

Myung, I. J., Kim, C., & Levy, W. B. (1997). Context-dependent Recognition in a Self-organizing Recurrent Network. In M. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Meeting of the Cognitive Science Society,* pp 530-535 Mahwah, NJ: Lawrence Erlbaum.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18,* 645-659.

한국심리학회지 : 실험 및 인지
The Korean Journal of Experimental and Cognitive Psychology
2002, Vol. 14, No. 4, 309 - 319

Brain and Cognitive Modelling: Text categorization using extended Latent Semantic Analysis

Cheongtag Kim Tahyun Lee

Department of Psychology, Seoul National University

The purpose of this study is to design a technique of text categorization which utilizes the meanings captured by a statistical analysis of a corpus. The meaning of a word was defined by a frequency pattern of texts which co-occurs with the word. The space of meanings is mathematically constructed from the results of analyses of the contingency table of words and documents, and texts were represented in the space. A text categorization system was designed by utilizing information about the associations among texts in the space. The statistical technique which was used in the analysis was Singular Value Decomposition and axis rotation, and the performance of text categorization of the system is similar to human beings'. The performance was enhanced when axis rotation technique was applied.

Key words Latent Semantic Analysis, text categorization, context, neural network

원 고 접 수 : 2002. 12. 1
최종게재결정 : 2002. 12. 22