2001년 한국정보처리학회 춘계 학술발표논문집 제8권 제1호

클러스터링을 이용한 텍스트 특성 인식

이근무*
* 경주대학교 컴퓨터공학과

Text Characteristic Recognition Using by Clustering

Rhee KunMoo
*Dept of Computer Science, Kyongju University

요약

. 텍스트 특성을 인식하는 방법적 접근은 텍스트의 기본적 특성을 이용하는 것에서부터 다변량 기법까지 다양한 방법이 제안되고 이용되고 있다. 이 논문에서는 이런 여러 기법들 중 클러스터링 기법을 이용하여 텍스트의 특성을 인식하고 그 인식능력의 효과성을 확인하고자 하였다. p 개의 변수로 구성된 N 개의 개체들은 p-차원 공간에 흩어진 N 개의 점으로 생각될 수 있으며 이들이 어떤 의미의 조밀성을 가지고 cluster를 이루고 있는지에 대한 정보는 자료의 구조를 이해하는데 매우 중요한 의미를 가지게 된다. 이런 결과들은 현재 우리학계의 도작사건논쟁, 인문학계 특히 고대사학과 민족 정체성에 대한 거대 담론들에 대한 여러 형태의 유용한 실증적 전거를 마련해주게 될 것이다.

1. 서론

텍스트 특성을 인식하는 방법적 접근은 기본적 특성에서부터 다변량 기법 까지 다양한 방법이 제안되고 이용되고 있다. 이 논문에서는 이런 여러 기법들중 클러스터링 기법을 이용하여 텍스트의 특성을 인식하고 그 인식능력의 효과성을 확인하고자 하였다. p 개의 변수로 구성된 N 개의 개체들은 p-차원 공간에 흩어진 N 개의 점으로 생각될 수 있으며 이들이 어떤 의미의 조밀성을 가지고 cluster를 이루고 있는지에 대한 정보는 자료의 구조를 이해하는데 매우 중요한 의미를 가지게 된다. 이러한 방법들은 그 원시자료에 대한 관찰과 조사가 정확히 진행된다나 현재 그 진위가 한국 고대사 연구의 초미의 관심이 되고 있는 화랑세기 등 고대사 저술의 진위를 판단하는 방법이 될 수 있으며 나아가 고미술품 등의 진위 판정에도 응용될 수 있을 것이다. 이런 방법론의 전통은 성서의 바울서한의 저자의 진위 논쟁, 세익스피어 작품의 진위논쟁 등 세계적 관심이 되는 것에서부터 정치적 사건 및 개인의 송사에까지 다양한 스펙트럼에서 논구될 수 있으며 이런 결과들은 현재 우리의 인문학계 특히 고대사학과 민족 정체성에 대한 거대 담론들에 대한 여러 형태의 유용한 실증적 전거를 마련해주게 될 것이다. 이 논문에

서는 화랑세기 환단고기등 상고 및 고대사의 여러 역사 텍스트의 진위를 결정하는데 이용될 수 있는 텍스트 특성을 인식하는 방법론으로써 clustering을 이용하고자한다. 텍스트의 진위를 확인하는 방법론의 전통은 성서의 바울서한의 저자의 진위 논쟁, 세익스피어 작품의 진위논쟁 등 세계적 관심이 되는 것에서부터 정치적 사건 및 개인의 송사에까지 다양한 스펙트럼에서 논구될 수 있다. 문헌의 작자에 대한 진위 판정은 지금까지 첫째 기록 내용에 대한 검토와 역사적 사실에 대한 고증, 둘째 원본 자료가 있을 경우 그 필적의 감정, 셋째 사용된 종이의 지질, 잉크 혹은 먹물 등의 화학적 분석에 의한 연대 추정에 의존하는 것이 보통이었다. 그러나 원저자의 원고가 남아있지 않지 않을 경우에는 필적감정이나 화학적 분석을 할 수 없거나 역사적 사실이나 고증을 할 수 없는 경우가 존재한다. 특히 고대 문헌의 경우에는 필사본으로서 전해오는 경우에는 필적감정이나 화학적 분석이 불가능하다. 또한 근년에 와서는 컴퓨터를 이용한 편집작업이 진행되면서 이러한 전통적인 방법으로는 문헌의 진위판단에 한계가 있다. 따라서 이러한 문제해결을 위한 새로운 과학적 방법이 필요하게 되었다. 여기에서 새롭게 대두된 방법론들이 한 문장의 길이, 단어의 문자수, 품사의 출현율, 특수한 말의 출현율 등의 문장의 수량적 특성을 조사하여 분석함으로써 저자의 문장스타일과 습관 등을 계량적으로 파악하여 문헌의 진위 여부를 판정하는 방법론이 제안되고 있다. 문헌의 스타일과 습관을 확인한다는 것은 개인에게 지문이 개인을 구별하는 증거가 되는 것처럼 개인의 문장의 특성 역시 진위 판단의 근거가 될 수 있다는 가정에 근거하고 있다.

이러한 계량적 방법에 의한 작자 진위에 대한 연구는 유럽의 경우 100여 년 이상의 전통을 가지고 있으나 우리나라에서는 연구자들의 식견으로는 아직 일천한 것이 사실이다. 현재의 고대사와 관련되어 그 진위가 의심받는 화랑세기와 같은 문제 저작들에 대한 진위판정에 효과적으로 이용할 수 있는 방법론과 시스템을 설계하고자 한다. 이 논문은 제1절, 제2 절에서는 clustring 기법을 제시하고 제3절에서는 관련연구를 살펴보고 제4절에서 결론을 맺는다.

2. clustering

clustering의 방법은 군집의 형태와 사용되는 상사성(또는 비상사성)의 척도와 연관되어 다양한 방법들이 있다. 일반적으로 고려되고 있는 변수가 세 개 이하인 경우에는 산점도등을 활용한 목측(目測)에 의해 군집관계를 파악하는 것도 바람직하나 변수의 수가 늘어나면 이러한 방법은 점차 어려워지고 특히 연구자의 주관적 판단이 중요한 역할을 하게 된다. 더욱이 서로 다른 군집방법들은 상당히 다른 결과를 보일 수도 있어 사용된 군집방법이 가지는 특성을 잘 이해하는 것이 무엇보다 더 실제 분석에 도움이 된다고 할 있다. 그럼 이제 일반적으로 많이 이용되는 방법들로서 계보적 군집방법과 최적분리 군집방법을 다루어 보자.

유사성(Similarity)과 거리(Distance)의 척도

(1) 표본평균벡터와 표본공분산행렬

① 자료행렬

p 개의 변수로 이루어진 n 개 개체(관측치)의 관측자료를 다음과 같이 표현하자.

$$X=\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ & \vdots & & \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}=\begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}$$

② 표본평균벡터 ($p\times 1$)

$$\overline{X}=\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}X_{i1} \\ \frac{1}{n}\sum_{i=1}^{n}X_{i2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}X_{ip} \end{pmatrix}$$

③ 표본공분산행렬 ($p\times p$)

p 개의 변수에 대한 표본공분산행렬은 다음과 같다.

$$S=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})'$$

(2) 거리의 척도

서로 다른 개체 $X_i=(X_{i1},X_{i2},\cdots,X_{ip})'$와 $X_j=(X_{j1},X_{j2},\cdots,X_{jp})'$ 사이의 거리(Distance)

$d_{ij}=d(X_i, X_j)$ 는 일반적으로 다음의 조건을 만족한다.

- $d_{ij}\geq 0$, $d_{ii}=0$
- $d_{ij}=d_{ji}$
- $d_{ik}+d_{jk}\geq d_{ij}$

위의 조건을 만족시키는 거리는 두 개체의 비유사성(Dissimilarity)의 척도이다. 두 개체 사이의 거리의 종류는 일반적으로 다음과 같다.

① 유클리드 거리

$$d_{ij}=\sqrt{(X_i-X_j)'(X_i-X_j)}$$

② Mahalanobis 거리

$$d_{ij}=(X_i-X_j)'S^{-1}(X_i-X_j)$$

③ Minkowski 거리

$$d_{ij}=[\sum_{k=1}^{p}|X_{ik}-X_{jk}|^m]^{1/m}$$

(3) 유사성의 척도

두 개체의 유사성(Similarity)

$s_{ij}=s(X_i, X_j)$

는 일반적으로 두 개체에 대한 변수들 사이의 상관계수를 많이 사용하며 그 식은 다음과 같다.

$$s_{ij}=\frac{\sum_{k=1}^{p}(X_{ik}-\overline{X}_{i\cdot})(X_{jk}-\overline{X}_{j\cdot})}{\sqrt{\sum_{k=1}^{p}(X_{ik}-\overline{X}_{i\cdot})^2}\sqrt{\sum_{k=1}^{p}(X_{jk}-\overline{X}_{j\cdot})^2}}$$

여기서,

$$\overline{X}_{i\cdot}=\frac{1}{p}\sum_{k=1}^{p}X_{ik}$$

(4) 거리와 유사성의 관계

① $d_{ij}1-s_{ij}$

② $d_{ij}=\sqrt{1-s_{ij}}$

③ $d_{ij}=\sqrt{2(1-s_{ij})}$

과 같은 방법으로 거리를 구할 수 있다.

3. 관련연구

1851년 런던대학 수학교수이고 논리대수의 창안자인 드 모르간(Augustes De Morgan)은 켐브리지의 목사 헤랄드에게 편지를 보내서 신약성경 중에서 바울서간 집필자에 대한 추정문제와 관련하여 각 서한에 대한 단어당 평균 문자수를 이용하여[1] 작자의 진위를 찾아낼 수가 있음을 밝혔다. 이 서한은 1882년 사후 그 부인에 의해 출판되었다.[2] 이를 본 오하이오대학의 물리학자 멘덴홀(T.C. Mendenhall)은 집필자를 추정하는데 모르간이 제안한 단어당 평균 문자수보다는 단어 길이의 범위(range)와 분포를 이용하는 것이 더욱 효과적이라는 제안을 하였다. 이 방법을 이용하여 멘덴홀은 1887년 SCIENCE 지에 동일한 집필자가 작성한 문장에 나타나는 단어길이의 도수분포를 WATT SPECTRUM이라 이름하고 이를 이용하여 집필자를 추정할 수 있는가를 검토하였다.

1901년 멘델홀은 Shakespeare작품의 진위논쟁에 뛰어 들어 그때까지 Shakespeare의 생애에 대해 불분명한 것이 많아 그의 작품이 동시대의 유명한 철학자이며 정치가인 Francis Bacon을 위시해서 여러 동시대 인물들이 쓴 것이라는 설이 퍼져 있었다. 멘덴홀은 Shakespeare와 Bacon의 작품에서 각각 40만 단어와 20 만 단어를 표집하여 단어의 길이의 분포를 측정하였다. 그 결과 Bacon은 1단어 당 3문자에 최빈값(mode)을, Shakespeare 는 4문자라는 것을 보여 주어 Bacon이 Shakespeare작품의 저자라는 설을 부정하였다.

그 이외에도 수학자 코시(A. L. Cauchy)는 평균, 중앙값, 사분위범위를, "그리스도에 대해서"(The Imitation Christ:De Imitatione Christi)의 저자를 찾는데 이용하였으며, 위에서 살펴 본 문장의 길이에 관한 척도에 대한 연구에서 나아가서 1944년 율(G. U. Yule)은 저자의 어휘량에 주목하여 그는 K 특성치를 제안하였다.[3]

작품 중에 단어의 나타나는 단어 Fi 가 Xi번 나타난다고 할 때

$$S1=\sum XiFi, S2=\sum X^2Fi,$$

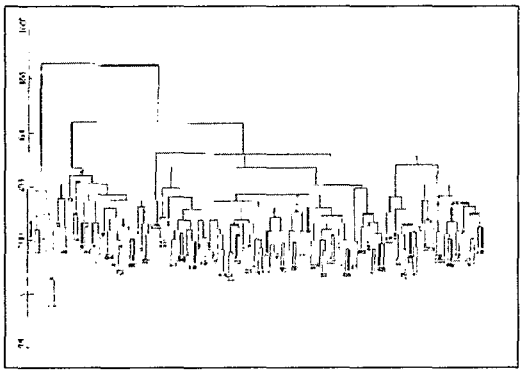
$$K=10^4((S2-S1)/S1^2)$$

1986년 藤泥偉作은 율의 데이타를 정준 상관을 이용하여 율의 연구결과를 지지하였다.[4]

이외에도 구약의 이사야서에 대한 진위 논의 역시 많은 시사점을 얻을 수 있다. 이사야서는 여러 사실적 의문으로 1970년대까지만 해도 39장을 전후로 전반부와 후반부가 다른 작자에 의해 쓰여진 것으로 이해되었다. 그러나 1973 년 Adams와 Rencher는

헤브라이어의 접두어의 사용비율을 이사야서에서 조사한 결과를 분석하여 이사야서의 전후반부 간에는 접두어의 사용율에서 높은 상관관계를 보이고 있어 이사야서는 1인에 의한 저작일 가능성이 높다는 연구를 발표하였다.[5] 그러나 1994년 村上征勝은 아래 이들의 방법론을 다른 구약성경에 대한 무선표집을 통해 상관분석을 한 결과 접두사를 이용한 상관분석의 결과는 표집한 모든 구약의 부분들에서 높은 상관을 보여 이를 통해 구약전체가 한사람의 집필자에 의한 저작으로 해석할 여지가 있음을 보여 Adams와 Rencher의 연구에 회의를 표시하고 이사야서 집필자에 대한 추정에 새로운 문제를 제기하였다.[6]

중국의 경우에도 중국의 대표적인 고대소설인 홍루몽(紅樓夢)이라는 소설의 경우 작품의 일부가 다른 사람에 의해 쓰여졌을 것이라는 논쟁이 계속 되어 오고 있었다. 1987년 夏旦大學 李賢平은 <그림 1>와 같이 120장의 각 장에 출현하는 47개의 虛字를 이용하여 주성분분석, 정준상관분석, clustering(clustering analysis)을 실시하였다. 그 결과 다음과 같이 결론을 맺었다. "홍루몽의 80회까지는 조설근이 내려오는 "石頭記"를 수정하여 쓴 것이고 그 이후는 120 회까지는 조설근의 자신의 초기소설 "風月寶玉"을 새롭게 추가하여 쓴 것을, 그 동



<그림1 > 홍루몽의 허사에 대한 clustering 결과

료들이 수정 가필한 것이라는 결론이었다.[7]

4. 결론

본 연구는 최근 역사학계의 중심논쟁으로 되고 있는 화랑세기 등의 저작 text에 대한한 진위를 밝히기 위한 방법론으로써 clustering 기법을 제안하고자 하였다. 이를 통하여 우리 역사의 지평과 전망을 다시 살펴보는 계기를 마련할 수 있을 것이다. 저작의 진위를 확인하는 연구에는 문학, 언어학, 통계학, 문헌정보학, 컴퓨터공학 등 다양한 학문분야에 의한 다학문적 접근을 필요로 한다.

이러한 연구의 성공을 위해서는 관련분야 연구자가 서로 지식이 부족한 부분에 대한 보완을 하면서 조직적 다면적 접근을 통한 광범위한 시야를 가지고 연구를 진행하여야 한다. 이를 위해서는 우선 각분야 연구자가 서로의 의견을 나눌 수 있는 기회를 만들어 다른 분야의 연구에 관한 지식을 이해하는 동시에 공통의 문제 의식을 가지고 공동 연구를 진행할 수 있는 체제를 확립하는 노력이 기울여야 할 것이다.

참고문헌

[1] R. D. Lord, Studies in the history of probability and statistics VIII. De Morgan and the statistical study of literature style. ***Biometrika,45,*** 1958

[2] S. E De Morgan, Memoir of Augustus de Morgan by his wife Sophie Elizabeth De Morgan with selection from his letters. Longman,Green and Co.,1882.

[3] G. U. Yule, The statistical study of Literacy Vocabulary . Cambridge University press, 1944

[4] 藤迟偉作, 未知の著作權の推定について。BASIC 數學 2 月号,1986

[5] L. L. Adams and A. C. Rencher, The popular critical view of the Isaiah problem in light of statistical style analysis. ***Computer Studies in the Humanities and Verbal Behavior*** . 7(3-4),1973.

[6] 村上征勝, 眞贋の科學, 朝倉書店, 1994.

[7] 李賢平, 紅樓夢成書新說, 夏旦大學,1987年 第5期