클러스터 分析과 情報檢索에 있어서 클러스터 分析의 應用

金　成　赫
(캘리포니아대학 환경
디자인연구소)

〈차　례〉

Ⅰ. 序　論

클러스터(cluster) 分析과 이 方法을 여러 分野에 應用시키기 위한 研究가 지난 수십년간 활발히 진행되어 왔다. 비록 現代的인 클러스터分析이 生物學的

인 分類學에서 출발하였지만, 이 方法은 모든 分野의 데이터에 일반적으로 應用되고 있다.

"클러스터링 (clustering)" 이란 말은 일반적으로 그룹을 構成하고 있는 個體間의 類似値(similarity)에 根據를 두고 그 個體들로 하여금 작은 그룹을 形成하게 하는 여러가지 方法들을 말한다.[1] 이 方法은 科學的인 硏究의 道具로서 利用되어 왔다.

클러스터分析의 가장 重要한 機能은 카테고리(category)에 관한 假定을 만드는 것이다. 그렇기 때문에 生物學에서 이 方法을 利用하고 있는 주된 理由는 動物이나 植物의 分類體系를 설정하기 위한 것이다. 또한, 우리는 이 方法을 통해 많은 量의 데이터를 비교적 밀집된 작은 데이터群으로 形成하는데 應用할 수가 있다.

클러스터링은 學問의 分野마다 다르게 불리어지고 있다.[2] 數値分類(numerical taxonomy)는 生物學者, 植物學者 및 生態學者들 사이에 클러스터링 代身으로 쓰여지고 있으며, 어떤 社會學者들은 類型學(typology)이라고도 한다. 人工頭腦學(cybernetics), 電氣工學 및 컴퓨터科學(computer science)에서 쓰여지고 있는 "learning without teacher"나 "unsupervised learning"은 클러스터링을 意味하고 있다. 情報檢索 및 言語學에서는 "clumping" 혹은 "automatic classification" 이라고 불리기도 한다. 地理學者들은 "regionalization"이라고도 하며, 그래프 설계자나 회로 설계자들 사이에 쓰여지는 "partition"은 클러스터의 집합을 일컫는다.

클러스터分析은 引用分析, 情報檢索 및 內容分析 등 情報科學의 여러분야에서 광범위하게 利用되어 왔다.

유(C.Y.Yu)[3]는 利用者의 질문식(query)에 기초를 둔 클러스터링 알고리즘을 提案하였다. 즉 系圖(tree)[4]의 뿌리에 모여있는 文獻들의 집단은

---

1) S.Miyamoto and K.Nakayama, "A Technique of Two-stage Clustering Applied to Environmental and Civil Engineering and Related Methods of Citation Analysis," *JASIS*, vol.34, no.3, 1983, pp. 192-201.

2) M.R.Anderberg, *Cluster Analysis for Application*, New York: Academic Press, 1973,

3) Clement Y.Yu,"A Clustering Algorithm Based on User Queries," *JASIS*, vol.25, no.4, 1974, pp.216-218.

4) 그래프(graph)의 한 種類로서, 데이터 구조(data structure)에 많이 利用되고 있다.

간단한 질문시에 적합한 문헌들이고, 系圖의 윗쪽에 모여있는 문헌들의 집단은 복잡한 문헌들이 되게끔 系圖構造의 클러스터링을 만들었다. 케슬러(M·H·Kessler)[5]는 문헌 사이의 類似値를 측정하기 위하여 書誌結合을 提案하였다. 즉, 두개의 문헌이 n개의 같은 문헌을 引用하였을 때 그 두 문헌 사이에는 類似値가 n이라고 하였다. 딜론(H.Dillon)[6]은 문헌의 內容을 分析하기 위하여 이 方法을 利用하였으며, 고틀리브(C.C.Gotlieb)와 쿠마르(S.Kumar)[7]는 索引語를 재구성하기 위하여 그래프理論을 利用한 클러스터링을 應用하였다.

그러나, 이러한 硏究들은 클러스터分析에 관한 理論的인 接近方法을 示提하지 못하고 있다. 단지, 그들은 分析하고자 하는 데이터에 적합한 알고리즘을 개발하고, 그 알고리즘과 다른 알고리즘을 시스템 효율[8] 및 클러스터 構成에 소요되는 시간만을 비교하고 있을 뿐이다.

이 글에서 우리는 일반적인 클러스터分析의 方法論과 이 클러스터分析이 情報檢索시스템 설계에 어떻게 應用되고 있는가를 검토해 볼 것이다.

또한 이 글에서 클러스터링, 그리고 키워드(keyword)와 用語는 같은 의미로 쓰이고 있다.

Ⅱ. 클러스터分析의 方法論

클러스터分析을 應用하는데 있어 가장 어려운 점은 클러스터分析이 어떤 確率이나 面積따위를 계산해내는 하나의 公式과 같은 것이 아니라 應用統計學의 여러 技法과 方法을 綜合的으로 利用해야 한다는 點에 있다. 실제로, 데이터를 갖고 클러스터를 構成하는 과정에서 우리는 어떤 技法을 사용해야 되는가를 直觀的으로 결정하기도 한다. 그렇기 때문에, 여러개의 클러스터分析技法 중에서 하나를 선택하는 기준 등을 나타내는 일반적인 틀(framework)이 必要하다.

---

5) M.H.Kessler,"Bibliographic Coupling between Scientific Paper,". *American Documentation*, vol.14,no.1, 1963, pp.10-25.

6) H.Dillon, "The Use of Clustering Techniques in the Analysis of Judical Mose," *Information Science*,1975.8, pp.10-25.

7) C.C. Gotlieb and S.Kumar, "Semantic Clustering of Index Terms," *JACM*, vol.15, no.4, 1968, pp.495-513.

8) 재현율(recall)과 적합률(precision)을 測定하여 비교하는 방법을 말한다.

1. 데이터의 特性 및 選擇

統計學에서 屬性(attribute)이란 말은 데이터의 特性을 나타내는데 利用되고 있다. 이 特性은 二進狀態에서는 특정한 키워드가 문헌에 나타나 있는가 혹은 나타나 있지 않는가, 즉 個體가 소유하고 있나, 혹은 소유하고 있지 않나를 두 상태(yes-no, 1-0)로 表示하는 方法이고, 正常的인 상태에서는 軍人의 서열과 그의 눈 색깔과 같이 各各의 個體는 오직 하나의 상태에만 속하는 유한한 상태의 집합[9]일 수 있다. 二進狀態는 情報檢索시스템에 널리 使用되기 때문에 뒤에서 자세히 설명될 것이다.

클러스터分析의 對象은 사람, 動物, 문헌, 用語 등일 수가 있다. 分析對象의 選擇에 있어 우리는 다음 2가지를 고려해야 할 것이다.

첫째로, 표본이 分析對象의 全體인 경우이다. 이 경우 分析目的은 주어진 데이터의 分類體系를 發見하는 것이다. 둘째로, 표본이 分析對象의 일부분인 경우이다. 클러스터分析은 두번째 경우에 더 많이 쓰여지고 있으며, 첫번째 경우에 클러스터分析을 適用시키는 것보다도 두번째 경우가 더 복잡하다.

統計學的인 方法은 無作爲 選擇原則(principle of random selection)과 獨立的 選擇原則(principle of independent selection)을 必要로 한다. 無作爲選擇이란 말은 모든 데이터는 표본으로 똑같이 選擇될 수 있다는 것을 意味하며, 獨立的 選擇이란 말은 어떤 특정 데이터의 選擇은 다른 데이터 선택에 影響을 미치지 않는다는 것을 말한다.

2. 變數의 選擇 및 스케일의 單一化

모든 데이터는 데이터의 特性 및 屬性에 의해 一致되게 記述하여야 한다. 이 特性 및 屬性들이 클러스터分析의 變數들이다. 클러스터分析에 있어서 變數의 選擇은 分析의 最終目標에 중요한 영향을 미치고 있다.

分析하고자 하는 모든 데이터가 거의 같은 變數를 갖고 있으면 클러스터形成이 밀집되어 있는 반면에, 데이터 사이의 變數에 차이가 있으면 클러스터形成이

---

9) 예를 들어 〈소위, 중위, 대위〉, 〈검정색, 갈색, 붉은색〉과 같이 個個의 個體는 두 집합의 원소 중 어느 한곳에만 속한다.

흩어지는 경향이 있다.

情報檢索시스템에 속해있는 문헌들은, 그 문헌들의 主題를 나타내 주는 用語들로 索引될 수가 있다. 이 用語들의 존재여부에 따라 클러스터의 모양이 다르게 나타날 수가 있다. 다시 말해서, 비슷한 用語들을 갖고 있는 문헌들은 같은 클러스터에 속하게 되며, 클러스터모양이 밀집되어질 것이고, 반대로 서로 다른 用語들을 갖고 있는 문헌들은 同一클러스터에 속하지 않고 클러스터모양도 흩어지게 된다. 그렇기 때문에 각 문헌에 나타나는 用語들을 變數라 할 수가 있다. 또한 索引時 혹은 情報檢索時 加重値用語들을 使用하는 경우와 非加重値用語를 使用하는 경우도 클러스터構成에 차이를 가져올 수 있기 때문에 이것도 하나의 變數로 볼 수가 있다.

架空的이 아닌 실제 데이터의 일반적인 문제점은 變數間의 同質性의 缺如이다. 다른 스케일을 갖고 있는 變數사이의 밀접관계를 測定하는 것은 어려운 문제점을 일으킨다. 變數사이에 다른 스케일을 갖고 있으면 정확한 클러스터를 形成하기가 어렵다. 그렇기 때문에 클러스터分析을 하기에 앞서 이용될 變數의 스케일을 하나로 統一시켜야 한다. 예를 들어, 한 變數에 그램과 톤이 使用되고 있으면, 올바른 클러스터를 形成할 수 없게끔 된다.

3. 무엇을 클러스터할 것인가.

클러스터의 對象은 데이터와 그 데이터가 갖고 있는 變數이다. 예를들어 情報檢索시스템에서 문헌들을 클러스터할 것인지, 또는 用語들을 클러스터할 것인지를 결정해야 한다. 일반적으로 變數의 클러스터構成이 데이터의 클러스터構成보다 복잡하다. 그렇기 때문에 데이터에 나타난 變數의 相互關係에 의해 變數를 클러스터하는 것이 바람직하다. 데이터와 變數를 同時에 클러스터하는 方法도 있다.

變數의 클러스터分析은 要素(factor)分析과 많은 점에서 類似하다. 비록 두 分析方法이 서로 다른 技法을 使用하고 있지만, 궁극적으로는 變數 사이의 관계를 밝히려고 하는 것이다. 技法의 차이 때문에, 만약 두 分析方法을 함께 利用하면 두 方法이 서로 補完하여 상당히 다양한 결과를 기대할 수가 있다.

클러스터分析과 要素分析은 다음과 같은 類似點을 갖고 있다. 要素分析은 變

數의 集團을 적은 數의 假定的인 變數로 나타내기 위하여 여러가지 統計學的인 技法을 使用하는 分析方法으로서 密接關係를 測定하기 위하여 相關係數를 채택하고 있으며, 이를 利用하여 相關行列을 만든다.[10] 相關行列은 하나의 變數와 다른 모든 變數와의 相互密接關係를 나타내고 있으며, 클러스터分析의 類似行列(similarity matrix)과 同一하다. 클러스터分析과 要素分析은 類似行列에 基礎를 두고 클러스터를 찾는 體系的인 方法들이다. 이 두 分析方法은 데이터分析에 널리 利用되고 있다.

보르코(H.Borko)와 베르니크(M.Bernick)[11]는 문헌의 그룹을 찾기 위하여 要素分析을 利用하였다.

4. 密接關係測定(similarity measure)

모든 클러스터分析 方法들은 클러스터되어지는 모든 對象間의 密接關係를 測定하는 方法이 있어야 한다.

예를 들어, 문헌들(D1, D2, D3)을 클러스터한다고 할 경우에 D1과 D2, D3, D2와 D1, D3 그리고 D3와 D1, D2 사이의 密接關係를 數値로서 나타내야 한다. 데이터를 클러스터할 경우에 이들 個體間의 밀접관계는 거리로서 測定된다. 이 경우, 個體를 平面위의 한 點으로 생각하면, 이들 個體間의 거리는 다음 조건을 만족시켜야 한다.[12] 다시 말해서 E는 測定空間을 나타내는 全體集合이라하고, X, Y, Z는 E속에 있는 어느 한 點이라 하자. 이 때 거리함수(distance function) D는 아래 조건을 충족시켜야 한다.

① $D(X, Y)=0$ iff[13] $X=Y$

② $D(X, Y)\geqq 0$ E안에 있는 모든 X, Y

③ $D(X, Y)=D(Y, X)$ E안에 있는 모든 X, Y

④ $D(X, Y)\leqq D(X, Z)+D(Y, Z)$ E안에 있는 모든 X, Y, Z

---

10) Kim Jae-On and Charles W. Mueller, *Introduction to Factor Analysis-What It Is and How to Do It*, (London : Sage University Press, 1978).

11) H. Borko and M. Bernick, "Automatic Document Classification Ⅱ - Additional Experiments," *JACM*, 1963.10, pp.152-162.

12) S.C. Johnson, "Hierarchical Clustering Schemes," *Psychometrika*, vol.32, no.3, 1967, pp.241-254.

13) if and if only.

첫 번째 것은 그 자신의 거리를 말한다. 다시 말해서 어느 두 점의 거리가 0으로 나타나면 그 두 점은 同一한 점을 의미한다. 두 번째는 마이너스 거리를 인정하지 않고 있으며, 세 번째는 두 점 X와 Y의 거리는 Y와 X의 거리와 같다. 즉, 對稱을 의미하고, 네 번째는 삼각형에서 한 변의 길이는 다른 두 변의 길이의 合보다 길 수가 없다는 것을 말한다.

變數를 클러스터할 경우에는 주로 相關關係를 利用하고 있다.

密接關係를 測定하는 方法은 매우 다양하다. 수많은 公式이 提案되었고, 그 중에서 어떤 것은 실제적으로 適用하기가 容易하지만, 어떤 것은 상당히 어려운 것도 있다. 그렇기 때문에 密接關係를 測定하는 公式의 選擇은 클러스터할 데이터의 特性에 의해 결정되어야 한다.

5. 클러스터링 構成基準

일반적으로 우리는 "클러스터"란 用語를 定義하지도 않은 채, 마치 幾何學에서의 點과 같이 基本的인 개념으로 취급하고 있다. 이와 같은 개념은 理論的으로 接近할 때에는 적합하지만, 실제적인 데이터를 利用하여 클러스터를 發見할 時에는 명확하게 그 범위가 한정되어야 한다.

클러스터 構成基準의 選擇은 클러스터를 定義하는 것과 같다고 볼 수 있다. 그러므로, 구성기준을 알고리즘의 수행과 기준의 明文化를 통해 설정한다는 것은 매우 중요한 일이라 하겠다.

지금까지 수 많은 서로 다른 구성기준이 제안되었고, 이용되어 왔으나, 완벽한 하나의 기준이 있는 것은 아니다.

따라서 클러스터構成時 하나의 基準을 使用하는 것보다는 클러스터 구조의 여러 면을 밝히기 위해 여러 개의 기준을 사용하는 것이 효과적일 것이다.

6. 알고리즘 선택과 컴퓨터 이용

앞에서 言及한 "데이터 및 變數의 선택", "무엇을 클러스터할 것인가", "밀접관계 특정공식", "구성기준" 등이 정하여졌다해도 실제로 클러스터를 구성하기 위해서는 문제점이 있다. 즉, 클러스터를 構成하는 알고리즘의 選擇이다. 한 그룹의 데이터를 가지고 여러가지 다른 알고리즘을 사용했을 경우, 궁극적으

로 똑같은 결과를 가져온다할지라도 알고리즘의 선택은 문제가 되는 것이다.

왜냐하면, 클러스터分析은 컴퓨터를 利用하여야 하기 때문에 알고리즘을 코딩한 컴퓨터 프로그램을 分析해 볼 필요가 있다. 지금까지 개발된 프로그램만도 그 수를 헤아릴 수가 없다. 어떤 프로그램은 클러스터할 데이터의 수가 정해진 것도 있고, 어떤 것은 지나치게 컴퓨터의 主記憶容量을 요구하는 것도 있고, 또 클러스터를 構成하는데 시간이 지나치게 소요되는 것도 있다.

클러스터分析을 利用하는데 있어 실제적이고 중요한 문제는 클러스터의 數를 결정하는 것이다. 어떤 알고리즘은 처음부터 일정한 수의 클러스터를 준 다음, 알고리즘이 진행함에 따라 類似値와 構成基準에 의해 그 수를 변화시키는 것도 있다. 그러나, 클러스터의 숫자를 정한다는 것은 어려운 일임에는 틀림이 없다.

7. 結果의 解釋

이와 같이 하여 클러스터들이 形成되었으면, 그 결과를 올바르게 해석하여야 할 것이다. 클러스터分析은 단순히 클러스터들의 集合만 구성하는 것이 아니라 分析하고자 하는 데이터를 理解하게 하고 그 特性을 알 수 있게 하는 것이다. 더우기 情報檢索시스템에서 문헌들을 클러스터하였을 경우, 그 결과는 같은 主題를 갖고 있는 문헌들이 한 클러스터에 모여있을 뿐만 아니라, 검색시간도 단축시켜주고, 나아가 利用者의 정보검색에 대한 만족도도 향상시킬 수 있게 되는것이다.

또한 클러스터分析의 결과는 分析者로 하여금 分析데이터를 통해 豫測 및 假定을 할 수 있게끔 도와주는 역할도 한다. "豫測은 여러 측면에서 생각해 볼 수 있다. 가장 간단하게는 分析者가 클러스터를 통해 데이터에 나타나 있지 않은 데이터의 特性을 發見할 수도 있고, 또 그 데이터와 類似한 데이터에도 利用할 수 있다. 한 단계 더 나아가 클러스터 결과의 分析을 통해 分析者는 分析데이터의 일반적인 假定(假說)을 설정하는데 도움을 줄 수도 있는 것이다.

클러스터分析者는 결과의 해석 및 중요도를 판단하는데 능동적으로 대처하여야 한다. 이 과정은 主觀的이고 直觀的이어야 한다.

클러스터 결과를 해석하고 중요도를 판단하기 위하여, 分析者는 다음과 같은 方法을 利用할 수 있다.[14]

---

14) A.O.Gordon, *Classification*, (London : Chapman and Hall, 1981).

1) 假說 및 檢定

- 分析데이터가 전혀 클러스터를 구성하지 않은 경우를 위한 歸無假說 測定
- 특정한 클러스터 構造를 發見하기 위한 測定

2) 시뮬레이션硏究

分析者는 이미 알려져 있는 데이터의 클러스터들을 통해 자기가 구성한 클러스터들과 여러 평가기준으로 검토해 볼 필요가 있다.

3) 比 較 硏 究

- 다른 類似値 公式을 利用하여 그 결과를 검토해 보기도 하고, 또 다른 變數를 使用해 보기도 한다.
- 다른 알고리즘을 적용시켜 보기도 한다.

Ⅲ. 클러스터링 方法

일반적으로 클러스터링構成方法은 階層클러스터링과 非階層클러스터링方法으로 구분하고 있다.

階層클러스터링方法은 클러스터링形成에 관한 자세한 순서를 系圖를 통해 보여주기 때문에 이해하기가 쉬운 반면에 非階層클러스터링方法은 많은 양의 데이터를 취급할 수가 있다.

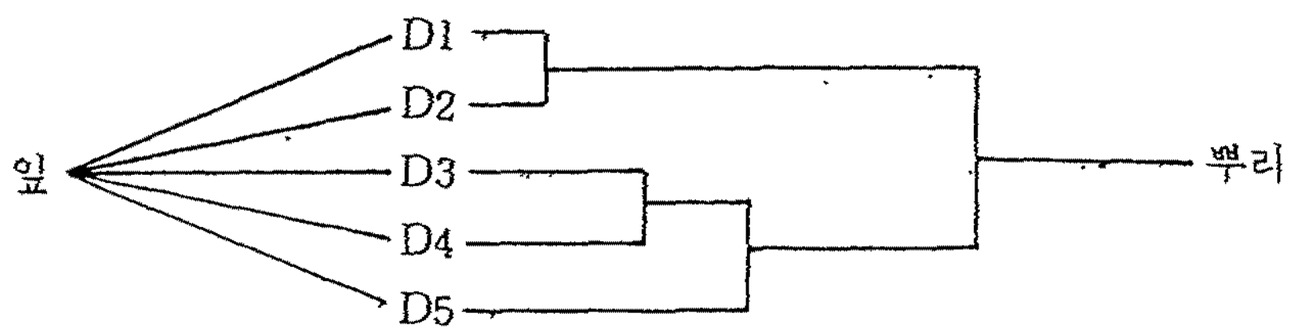
1. 階層 클러스터링方法

계층클러스터링方法은 데이터를 여러 등급으로 요약하는데 가장 널리 利用되고 있다. 類似値는 데이터 個體間의 관계를 나타내주는 類似行列을 構成하는데 이용된다. 즉, 계층클러스터링은 이 행렬(圖 1)을 이용하여 個體間의 관계를 보

〈圖 1〉 類 似 行 列

	D1	D2	D3	D4
D5	.21	.31	.41	.50
D4	.35	.29	.65	
D3	.30	.41		
D2	.85			

〈圖 2〉 系 圖



여주는 系圖를 만드는 것이다(〈圖 2〉).

:예를들어 5개의 문헌을 클러스터링한다고 할 경우, 類似行列은 〈圖 1〉과 같이, 系圖는 〈圖 2〉와 같이 나타낼 수 있다. 〈圖 2〉에서 왼쪽에 있는 잎들은 個個의 문헌(D1, D2, D3, D4, D5)을 나타내고 있으며, 뿌리는 문헌전체의 集合을 보여주고 있다. 잎으로부터 뿌리에 이르는 과정은 〈表 1〉과 같이 나타낼 수 있다.

이와같이 잎으로부터 시작하여 뿌리에 이르면서 系圖를 구성하는 階層클러스터링方法은 또한 凝集(agglomerative)方法이라고도 불리어지고 있다. 여기서 注目해야될 것은, 이들 클러스터들은 포개진다는 것이다. 즉, 잎으로부터 뿌리를 향해 뭉쳐지는 입장에서 볼 때, 두 개의 문헌이 모여서 하나의 그룹으로 이루어지고, 두 문헌은 영원히 합쳐지게 되는 것이며, 다음 단계를 위해 블럭(block)을 形成하게 된다. 반면에 갈라지는 입장으로 보면, 문헌의 작은 그룹들이 二部分으로 나뉘어지는 것은 영원하게 분리되는 것으로, 남아있는 다른 문헌과는 관계없이 독립적으로 취급되는 것이다.

이것이 바로 階層클러스터링의 長點 및 短點이다. 즉, 처음부터 그룹을 形成해 나가기 때문에 검토해야 할 데이터가 상당히 줄어드는 반면에 初期의 실수나 나중 기회를 利用하는 것이 排除된다는 것이다. 지금까지 提案된 階層 클러스터

〈表 1〉 클러스터 形成過程

클 러 스 터 수	클 러 스 터
5	(D1), (D2), (D3), (D4), (D5)
4	(D1, D2), (D3), (D4), (D5)
3	(D1, D2), (D3, D4), (D5)
2	(D1, D2), (D3, D4, D5)
1	(D1, D2, D3, D4, D5)

링方法은 너무나 많기 때문에 일일이 열거할 수가 없지만 다음과 같이 분류해 볼 수 있다.

(1) 凝集(agglomerative)方法

링키지(linkage) 方法, 센트로이드(centroid)方法 및 誤差의 自乘合 또는 分散方法들이 여기에 속한다. 링키지方法은 다시 단일링키지, 完全링키지, 平均링키지로 區分할 수 있다.

단일링키지는 두 데이터, 혹은 두 클러스터 사이의 가장 가까운 거리를 凝似値로 利用하여 클러스터를 構成하는 方法이고, 완전링키지는 반대로 가장 먼 거리를 類似値로 利用한다. 平均 링키지는 클러스터간의 평균거리를 類似値로 利用하는 方法으로서, 평균거리는 算術平均에 의해 산출된다.

센트로이드方法은 두 클러스터間의 센트로이드를[15] 계산하여 클러스터를 구성하는 方法으로, 이 경우 두 클러스터 혹은 個體間의 類似値가 클러스터 혹은 個體 i와 j의 센트로이드 사이에 있지 않으면 전혀 엉뚱한 클러스터가 구성될수도 있다.

이외에도 워드(J.H.Ward)[16]의 誤差自乘合의 目的函數를 利用한 方法과 分散을 利用한 것이 있다.

(2) 分割 (divisive)方法

이 方法은 系圖의 뿌리로부터 시작하여 잎을 향해 뻗는 階層系圖를 構成한다. 다시 말해서, 全體데이터 집단을 작은 그룹으로 나누는 것에 基礎를 두고 있다.

2. 非階層클러스터링方法

클러스터링方法과는 달리 처음부터 k개의 클러스터를 갖고 시작하는데, k는 分析에 앞서 결정되거나 혹은 클러스터링과정에서 결정되기도 한다. 여기에 속하는 대부분의 方法들은 分析하고자 하는 데이터를 처음부터 몇 개의 클러스터

---

15) center of gravity 라고도 하며, cluster의 重心을 말함.

16) J.H.Ward Jr. "Hierarchical Grouping to Optimise an Objective Function," *J. of American Statistics Association*, vol.58, no.301, 1963, pp.236-244.

로 나누어 놓고, 더 나은 클러스터를 구성하기 위하여 類似値에 근거를 두고, 각 클러스터에 속한 데이터를 교환하는 것이다. 여기에 사용되는 알고리즘들은 어떻게 하면 더 나은 클러스터를 構成할 수 있나, 어떤 方法을 利用하여야 초기 클러스터들을 개선할 수 있는가에 초점을 두고 있다.

비계층클러스터링方法은 類似値行列을 계산하여 컴퓨터에 기억시킬 必要가 없기 때문에 階層클러스터링보다 많은 양의 데이터를 취급할 수가 있다.[17] 왜냐하면, 클러스터할 데이터는 順次的으로 처리되며, 필요한 경우, 데이터는 磁氣테이프나 디스크로부터 읽혀지게 된다.

列擧分割(enumeration of partition)方法은 클러스터하고자 하는 데이터 집단을 하나 하나 열거해 가면서, 적절한 指標(index)를 利用하여 가장 적절한 클러스터링을 選擇해 가는 것이다. n개의 데이터를 서로 다른 k그룹으로 분할할 경우, 그 분할숫자가 너무 커지기 때문에 이 方法은 非實用的이다.

近接센트로이드分割(nearest centroid sorting)方法은 클러스터되어질 데이터의 주위에서 클러스터核으로 利用될 發芽點(seed point)들을 選擇하여, 클러스터構成에 利用하는 것이다. 이 發芽點들은 클러스터들의 센트로이드로서 계산되고, 각 데이터는 자신의 센트로이드와 가장 가까운 發芽點을 가진 클러스터에 속하게 되며, 이 과정을 되풀이함에 의해 클러스터가 構成되는 것이다. 發芽點은 매번 갱신될 수도 있고 또는 한 週期가 끝난 후에 갱신될 수도 있다.

이외에도 分散共分散을 利用한 方法이 있다.

Ⅳ. 클러스터링方法의 選擇 및 比較評價

지금까지 利用된 클러스터分析方法과 이에 使用된 類似値 公式이 너무나 많기 때문에 클러스터分析者는 자신이 利用할 方法을 選擇할 때 당황하게 된다. 選擇의 對象이 많다는 것은 그 중에서 어떤 特定한 方法도 자신이 分析하고자 하는 데이터에 적합한 것이 없다는 것을 意味하기도 한다.

---

17) M.R. Anderberg, *op. cit.*

分析者는 선택해야 할 方法들을 理論的 혹은 槪念的으로 檢討해 볼 수 있지만, 分析方法과 類似値 公式이 너무나 다양하기 때문에 이것만으로는 充分치 못하다. 어떤 주어진 데이터에 대해서 여러가지의 클러스터方法이 같은 결과를 낳을 수도 있는 반면에, 方法들 間의 차이 때문에 전혀 다른 결과를 가져올 수도 있다.

여러 클러스터링方法들 사이의 類似點과 相異點은 데이터의 特性을 이미 알고 있는 데이터를 클러스터링한 결과를 비교·분석하여 밝힐 수가 있고, 또 데이터 집단의 特性은 클러스터링構成의 속성을 알고 있는 方法들을 利用하여 클러스터링을 構成함으로써 밝혀 낼 수가 있다. 클러스터링方法들을 比較 評價하기 위한 基準은 다음과 같이 요약될 수 있다.[18]

(1) 外部的인 評價基準

클러스터方法의 外部的인 評價基準은 주어진 데이터에 대해서 여러 클러스터링方法을 利用하여 얻어진 클러스터모양들을 비교함으로써 評價될 수가 있다. 이들, 클러스터모양 사이의 類似程度 및 차이점은 다음과 같은 統計學的인 方法을 利用하여 測定된다.

- 서로 다른 方法을 利用하여 形成된 두 개의 클러스터모양에 나타난 데이터의 相互關係를 表示한 分割表를 利用한 測定方法으로서, 두 클러스터링 사이의 一致되는 부분을 백분율(%)로 나타낸다. 다시 말해서, 전체 데이터 중에서 두 클러스터링모양에 일치하는 데이터의 수를 전체 데이터 숫자로 나눈다.

  또 클러스터링 모양 사이의 차이를 χ^2 (chi − square)검증을 통해서 알아볼 수도 있다.
- 相關係數를 利用한 두 클러스터링의 類似程度 測定으로서, 두 클러스터링으로부터 類似行列을 만들어 相關係數를 계산하여 검토하는 方法이다.

(2) 內部的인 評價基準

주어진 데이터에 대해서 構成된 클러스터링 결과의 적절함을 測定하는 것인데,

---

18) Juan E. Mezzich and Herbert Solomon, *Taxonomy and Behavioral Science*, (New York: Academic Press, 1980).

이것은 소칼(R.R.Sokal)과 롤프(F.J.Rohlf)[19]가 소개한 " cophenetic 相關係數"에 의해 測定된다. 이것은 初期類似行列과 클러스터링 모양을 보고 구성한 類似行列에 나타난 데이터 사이의 相關關係를 계산한 積率相關係數이다.

이외에도 클러스터링方法을 評價하기 위해서는 分析하고자 하는 데이터가 갖고 있는 特性과 클러스터링 方法들이 갖고 있는 特性을 열거하여 比較해보는 것도 좋은 方法이다.

데이터가 갖고 있는 特性들은 분석할 데이터와 變數의 數, 클러스터할 對象의 選擇, 데이터가 갖고 있는 變數의 種類, 變數의 選擇 및 加重化, 클러스터의 數, 데이터 수집에 따른 문제점, 클러스터內의 데이터의 分布 등이 있고, 클러스터링 方法들의 特性은 결과의 형태(階層分類, 분할), 클러스터링에 소요되는 費用과 컴퓨터 기억용량, 類似値 測定公式과의 相互作用, 클러스터構造의 特性, 分析者의 直觀的인 決定에 대한 依存 등이 있다.

예를 들어, 어떤 클러스터링 分析方法은 적은 양의 데이터를 클러스터링 하는데 적합한 반면에 결과를 해석하기가 힘든 것도 있고, 어떤 것은 많은 양의 데이터에 적합한 반면에 클러스터링에 소요되는 시간과 비용이 많이 드는 것도 있다. 그렇기 때문에 뒤에 열거한 사항들을 도표로 만들어 比較하여, 分析하고자 하는 데이터에 적합한 方法을 選擇하여야 한다.

클러스터링 方法들의 評價는 클러스터 分析의 결과가 알고리즘의 수행과정 뿐만아니라 分析者의 分析目的, 해석능력에 의존하기 때문에 상당히 어렵다고 할 수 있다.

클러스터링方法들은 단순히 그 方法을 利用하여 얻은 결과가 옳은지, 또는 그른지만을 보고 평가될 수가 없는 것이다. 다시말해서, 클러스터링方法을 利用하여 얻어진 階層分類 또는 분할은 分析者에게 단순히 데이터를 物理的으로 재배열한 것을 보여주는 것이라기보다는 데이터의 特性 및 構造를 理解하게끔 도와주는 것이다.

---

19) R. R. Sokal and F. J. Rohlf, "The Comparison of Dendrogram by Objective Methods," *Taxon*, 1962.11, pp.33 ~ 40.

Ⅴ. 情報檢索시스템 설계에의 應用

지금까지 우리는 일반적인 클러스터分析의 方法論에 대하여 살펴보았다. 앞에서도 言及했듯이 이 方法들은 모든 分野에 適用시킬 수가 있는 것이다.

이제, 우리는 이러한 클러스터링의 일반적인 槪念을 念頭에 두고, 이 方法을 어떻게 情報檢索시스템 설계에 應用할 수 있는가에 대해 생각해보자.

클러스터링方法을 이 分野에 應用시키기 위한 硏究는 활발히 진행되어 왔으며 특히, 보르코(H.Borko) 와 베르니크(M.Berniek),[20] 살톤(G.Salton);[21] 그리고 반 리즈버겐(C.J.Van Rijsbergen)[22] 등은 이 方法을 통하여 檢索시스템의 효율을 개선할 수 있다고 주장하였다.

情報檢索시스템 설계에 있어서 클러스터링은 간단히 말해 主題가 類似한 문헌들을 같은 集團으로 모아놓는 것을 意味한다. 檢索시스템에서 문헌들은 主題를 나타내기 위해 用語들의 집합으로 구성되어 있다.

그렇기 때문에 이 用語들 사이의 類似値를 근거로 하여 클러스터링을 構成할 수가 있다.

檢索시스템에 이 方法을 利用하는 主目的은 檢索을 容易하게 하여줄 뿐만 아니라 檢索結果에 대한 利用者의 만족도를 개선시켜줄 수도 있다. 예를들어, 클러스터링을 利用한 情報檢索시스템을 생각해 보자. 이 시스템은 利用者의 질문식에 기초를 두어 類似한 문헌들끼리 클러스터링을 形成하고 있다. 즉, 유사한 用語들을 갖고 있는 문헌들은 自動的으로 同一한 集團 혹은 同一한 클러스터에 모여있기 때문에, 그리고 利用者의 질문식에 使用된 用語들과 類似한 用語들만을 갖고 있는 클러스터만을 조사하여 檢索하기 때문에 檢索時間을 단축시킬 뿐만 아니라 利用者가 願하는 情報를 정확히 검색해 낼 수 있는 것이다.

情報檢索시스템은 데이터검색과 문헌검색으로 區分될 수 있으며, 일반적으로 情報檢索이라함은 문헌검색을 말하고 있다. 문헌검색시스템에서 우리가 데이터

---

20) H. Borko and M. Bernick, *op. cit.*

21) G. Salton, *Automatic Information Organization and Retrieval*, (New York; McGraw-Hill, 1968)

22) C. J. Van Rijsbergen, "An Algorithm for Information Structuring and Retrieval, *The Computer Jr.*, vol.14, 1971, pp.407～412.

로 생각할 수 있는 것은 個個의 문헌, 索引用語 및 利用者의 질문식이다. 그렇기 때문에 문헌검색에 있어 클러스터링, 用語클러스터링 및 利用者의 質問式 클러스터링 3分野가 있다고 할 수 있다.

고틀리브(C.C.Gotlieb)와 쿠마르(S.Kumar)[23]는 用語들간의 語意를 밝혀 用語事典의 재구성에 관한 硏究에 이 方法을 이용하였고, 문헌클러스터링은 살톤(Salton)의 지도 아래 코넬大學에서 광범위하게 연구되어 왔다. 특히 이 硏究들은 정보검색시스템에서 클러스터 파일作成과 役割에 초점을 두었다. 최근에는 유(C.Y.Yu)가 利用者의 檢索結果에 대한 만족도를 개선하기 위하여 利用者의 질문식 클러스터링에 관한 연구를 하였다.

1. 類似値 測定

문헌검색에 있어서 다음과 같은 데이터들이 類似値 測定에 利用되고 있다.

① 한 쌍의 문헌벡터(vector)[25]

② 한 쌍의 索引用語 벡터

③ 문헌벡터와 분류벡터

④ 利用者 질문식 벡터와 문헌벡터

⑤ 利用者 질문식 벡터와 분류벡터

⑥ 한 쌍의 분류벡터

索引用語 벡터는 문헌 벡터를 逆으로 利用하여 만들 수가 있으며, 분류 벡터 혹은 利用者 질문식 벡터도 같은 方法으로 構成할 수 있다.

두 개의 문헌 D1과 D2의 유사치를 비교하는데 있어, 우리는 對稱關係를 기대하여야 한다. 즉, 두 문헌 사이의 유사치를 계산하기 위한 함수를 S라고 할 때, $S(X, Y) = S(Y, X)$는 당연한 것이다. 그러나 조건확률에 기초를 둔 유사

---

23) C.C. Gotlieb and S.Kumar, *op.cit.*

24) Clement Y. Yu, *op.cit.*

25) 情報檢索시스템에서 문헌, 用語, 利用者질문식, class는 다음과 같이 vector로 표시할 수 있다.

문헌 $Di = (d_1, d_2, \cdots dn)$ di는 i번째 用語를 말함.

用語 $Ti = (t_1, t_2, \cdots tm)$ ti는 用語 Ti를 색인어로 사용한 문헌을 말함.

利用者질문식 $Qi = (q_1, q_2, \cdots qk)$ qi는 질문식에 사용한 用語를 말함.

class $Ci = (c_1, c_2, \cdots ci)$ ci는 같은 class에 속해있는 문헌을 말함.

치 측정방법은 $P(X|T) \neq P(Y|X)$ 때문에 종종 非對稱일 경우가 있지만, 조건확률의 산술평균을 利用함으로써, 두 데이터 사이의 유사치를 對稱으로 만들 수가 있어 類似値 測定에 큰 장애요인은 아니라고 볼 수 있다.

類似値 測定을 위한 수 많은 公式 가운데, 情報檢索分野에 많이 쓰이는 것들은 아래와 같다.[26]

① $|X_T \cap Y_T|$

② $2\dfrac{|X_T \cap Y_T|}{|X_T| + |Y_T|}$ Dice의 公式

③ $\dfrac{|X_T \cap Y_T|}{|X_T \cup Y_T|}$ Jaccard의 公式

④ $\dfrac{|X_T \cap Y_T|}{|X_T|^{1/2} * |Y_T|^{1/2}}$ cosine의 公式

⑤ $\dfrac{|X_T \cap Y_T|}{Min(|X_T|, |Y_T|)}$ overlap의 公式

⑥ $\dfrac{|X_T \cap Y_T|}{|X_T| + |Y_T| - |X_T \cap Y_T|}$ tanimoto의 公式

여기에서 $|\ |$는 각 문헌이 갖고 있는 用語들의 數를 말한다.

첫번째 公式은 문헌에 使用된 用語의 數에 관계없이 利用될 수 있지만, 나머지 公式들은 색인에 사용된 用語들의 수를 고려해야 한다. 이러한 類似値 測定公式들은 두 벡터 사이의 類似程度를 결정하는데 利用되는 여러기준 즉, 두 벡터 사이에 일치하는 用語의 수, 각 벡터가 소유하고 있는 用語의 수, 혹은 이러한 索引用語들에 使用되는 가중치 등에 기초를 두어야 한다.

예를 들어 다음과 같은 二元關係를 갖고 있는 문헌-용어 행렬을 생각해보자.

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
D1	0	1	1	0	1	0	1	1	0	1
D2	1	0	1	1	1	0	1	0	1	1

26) C. J. Van Rijsbergen, *Information Retrieval*, (London: Butterworths, 1979).

문헌 D1은 T_2, T_3, T_5, T_7, T_8, T_{10}을 利用하여 索引되었고, 문헌 D2는 T_1, T_3, T_4, T_5, T_9, T_{10}으로 索引되었다고 가정하자. 여기서, 위에 열거한 유사치 공식을 이용하여 문헌 D1과 D2의 類似値를 測定하면 다음과 같다.

우선 $|D1| = 6$, $|D2| = 7$, $|D1 \cup D2| = 9$, $|D1 \cap D2| = 4$

① $S_T(D1, D2) = |D1 \cap D2| = 4$

② $S_T(D1, D2) = 2\dfrac{|D1 \cap D2|}{|D1| + |D2|} = 2 \cdot \dfrac{4}{6+7} = \dfrac{8}{13}$

③ $S_T(D1, D2) = \dfrac{|D1 \cap D2|}{|D1 \cup D2|} = \dfrac{4}{9}$

④ $S_T(D1, D2) = \dfrac{|D1 \cap D2|}{|D1|^{1/2} * |D2|^{1/2}} = \dfrac{4}{\sqrt{6} \times \sqrt{7}} = \dfrac{4}{\sqrt{42}}$

⑤ $S_T(D1, D2) = \dfrac{|D1 \cap D2|}{Min(|D1|, |D2|)} = \dfrac{4}{Min(6, 7)} = \dfrac{4}{6}$

⑥ $S_T(D1, D2) = \dfrac{|D1 \cap D2|}{|D1| + |D2| - |D1 \cap D2|} = \dfrac{4}{6+7-4} = \dfrac{4}{9}$

위에서 본 바와 같이 모든 측정공식이 서로 다른 측정치를 나타내고 있다. 그러면, 이들 公式 가운데 어떤 것을 使用하여야 가장 좋은 것일까?

일반적으로 클러스터링에 利用되고 있는 類似値 測定公式들은 다음과 같은 조건을 만족시키면 어떤 것을 利用하든지 클러스터링形成에 큰 차이가 없는 것같다.[27]

① 두 문헌 벡터가 공통의 用語들을 갖고 있지않을 경우 두 문헌 X와 Y의 類似値는 반드시 "0"으로 나타나야 한다.

② 두 문헌 벡터가 同一한 用語들을 갖고 있을 때, 두 문헌 X와 Y의 類似値는 最高値를 가져야 한다.

③ 類似値는 일정한 범위(－1과 1사이)를 갖고 있어야 한다.

클러스터링分析에 있어 類似値 測定公式의 선택은 클러스터링 構成에 있어 결

27) G. Salton, *op.cit.*

정적인 역할을 하지 못한다.[28] 다시말해서, 위와 같은 조건을 만족시키는 公式이면 클러스터링 分析에 利用할 수 있는 것이다. 그렇기 때문에 公式을 선택할 때 될 수 있는대로 간단한 公式을 선택하는 것이 바람직한 것같다.

2. 클러스터링을 情報檢索에 利用하기 위한 假定

클러스터링 分析을 문헌검색에 응용하려면, 먼저 다음과 같은 가정을 설정해야 한다.

① 클러스터들의 獨立

② 클러스터들의 重複

클러스터들의 獨立이란 한 그룹의 문헌들이 몇개의 클러스터들로 나뉘어졌을 경우, 이들 몇개의 클러스터중에서 어느 한 클러스터의 한 단계 더 깊은 분류는 기타 클러스터의 分類와 무관하다는 것을 의미하는 것이고, 클러스터들의 중복이란, 문헌검색시스템에서 문헌들은 2개 혹은 그 이상의 주제범위를 가질 수 있다는 것을 의미한다. 즉, 특정 문헌이 여러 개의 클러스터에 속할 수도 있음을 말하는 것이다.

이러한 가정하에서 다음과 같은 질문을 생각해 보자.

만약 클러스터링을 利用하여 문헌검색시스템을 개발하였을 경우, 어떤 특정 클러스터에 모여있는 문헌들은 서로 비슷한 여러 개의 질문식에 적합한 문헌들로만 구성되어 있을 수 있을까?

여기에 2개의 문헌검색시스템이 있다고 가정하자. 하나는 特定主題에 관한 시스템이고, 다른 하나는 모든 주제를 광범위하게 취급한 시스템이 있다. 즉, 前者는 거의 類似한 主題를 갖고 있는 문헌들로 구성되어 있는 반면에 後者는 전혀 다른 주제를 갖고 있는 문헌들로 구성되어 있다. 각각의 시스템에 위의 질문을 적용해 보자. 일반적으로 前者의 경우 위의 질문에 "예"라는 대답을 기대할 수 있지만 後者의 경우는 그렇지 못한 경향이 있는 것 같다.

반 리즈버겐은[29] 위의 질문에 대해 일단 "예"라고 가정을 해놓고, 시스템의 見

---

28) K. Sparck Jones, *Automatic Keyword Classification for Information Retrieval*, (London:Butterworths, 1971.).

29) C.J. Van Rijsbergen, *op.cit.*

本을 추출하여, 이것을 실험하였지만, 클러스터링을 利用한 모든 문헌검색시스템이 위 질문에 "예"라는 보장을 못하고 있다.

일단 이 실험에서 "예"라는 답이 나오면, 이것은 검색에 소요되는 시간을 단축시켜 줄 뿐만 아니라, 利用者의 정보요구에 대한 적합률도 증가시킬 수 있게 시스템을 설계할 수가 있다. 그렇기 때문에 클러스터링을 利用한 정보검색시스템은 유사한 주제를 갖고 있는 문헌들에 적합하다고 할 수 있다.

오늘날 문헌검색시스템은 학문의 細分과 더불어 특정 주제별로 나뉘어 개발되기 때문에 클러스터링의 利用이야말로 시스템의 효율을 개선할 수 있지않나 생각된다.

3. 알고리즘의 選擇

情報檢索에 利用되고 있는 클러스터링 알고리즘들이 너무나 많기 때문에, 이의 選擇에 앞서 다음과 같은 것을 생각해 보아야한다.[30]

첫째는 선택하고자 하는 알고리즘이 理論的인 근거를 갖고 있나를 살펴 보아야 하고, 둘째로 클러스터링의 효율 즉, 클러스터링에 소요되는 時間과 컴퓨터 기억용량을 얼마나 必要로 하는가를 검토하여야 한다.

理論的인 근거는 다음과 같은 기준을 만족시켜야 한다.

① 시스템에 있는 문헌들의 變動, 삭제, 수정, 첨가가 있어도 전체 클러스터링에 크게 영향을 미쳐서는 안된다.

② 문헌의 記述에 있어 작은 실수가 있더라도 전체 클러스터링에 크게 변화를 주어서는 안된다.

③ 최종적인 클러스터링構成이 초기 클러스터들과 독립적이어야 한다.

클러스터링의 효율은 클러스터링을 이용하였을 때 나타나는 검색효율을 말하기도 한다. 그러나 情報檢索에 利用되고 있는 클러스터링 알고리즘은 데이터의 種類에 따라 다르기 때문에, 클러스터하고자 하는 데이터에 적합한 알고리즘을 개발하는 것이 가장 바람직하다.

알고리즘을 개발할 때 유의할 점은 클러스터의 數, 각 클러스터의 최대 및 최소 크기, 檢索해 낼 수 있는 가중치의 부여, 클러스터간의 중복통제 등을 고려

30) Ibid.

해야 한다. 그러나 알고리즘의 선택은 기타 조건들을 選擇한 후에 결정되어야 한다.

4. 클러스터링 構成過程

이제 간단한 문헌 클러스터링 구성과정을 예로 들어 살펴보자.

문헌검색시스템에서 문헌들과 그 문헌들의 索引에 利用되는 用語들을 使用하여 二進形態의 文獻-用語行列을 구성할 수 있다. (圖 3)의 二進形態의 文獻-用語行列에서 매 行(row)은 文獻의 特性을, 매 列(column)은 用語의 특성을 나타내는 벡터이다. 다시 말해서, 매 行은 특정문헌이 소유하고 있는 用語들이 二進形態(0 또는 1)로 표시되어 있고, 매 列은 特定用語가 어떤 문헌의 索引에 利用되었는가를 역시 二進形態로 나타내고 있다. 이 行列을 이용하여 시스템에 있는 문헌들, 혹은 用語들 사이의 類似値에 근거를 둔 유사행렬을 만들 수 있다.

특정용어들이 특정문헌들과 서로 관련되어 있기 때문에 문헌들 혹은 용어들을 이용한 클러스터링은 똑같은 결과를 낳게된다. 그렇기 때문에, 本稿에서는 문헌들 사이의 類似値를 測定하여, 문헌-문헌 유사행렬을 만들었다. 문헌들간의 유사치를 측정하기 위해, 여기서는 자카드(Jaccard)의 公式을 利用하였다. 즉, 두 문헌 X와 Y의 유사치는

$$S_T(X,Y) = \frac{|X_T \cap Y_T|}{|X_T \cup Y_T|}$$

〈圖 3〉 二進形態文獻-用語行列(document-term matrix)

	T1	T2	T3	T4	T5	T6	T7	T8	T9
D1	1	0	1	0	0	1	1	0	1
D2	0	1	1	1	1	0	0	0	0
D3	0	0	1	1	0	0	1	0	0
D4	0	1	1	1	1	0	0	0	1
D5	0	0	0	1	0	1	1	1	1
D6	1	1	0	0	1	0	1	1	0
D7	0	1	1	1	1	0	1	0	1
D8	1	0	1	0	1	1	1	1	0
D9	0	0	0	0	1	0	1	1	1
D10	1	0	0	0	0	1	1	1	1

〈圖 4〉 문헌-문헌유사행렬(Document-document Similarity matrix)

(Jaccard의 공식을 이용하여 〈圖 3〉으로부터 계산해 냄)

①	D1	D2	D3	D4	D5	D6	D7	D8	D9
D10	.67	.00	.17	.11	.67	.43	.22	.58	.50
D 9	.29	.14	.14	.29	.50	.50	.43	.58	
D 8	.57	.25	.29	.22	.38	.58	.33		
D 7	.38	.67	.50	.83	.38	.38			
D 6	.25	.29	.14	.25	.25				
D 5	.43	.13	.33	.25					
D 4	.25	.80	.33						
D 3	.33	.40							
D 2	.13								

두 문헌 X와 Y의 유사치는 Y와 X의 유사치와 같다는 것은 당연한 가정이다. 즉, 유사치는 對稱($S_T(X,Y) = S_T(Y,X)$)을 이루어야 한다.

위의 유사치 공식을 利用하여 모든 문헌들간의 유사치를 계산하여 〈圖 4〉와 같은 문헌-문헌 유사행렬을 얻었다.

n개의 문헌을 클러스터링 한다고 할 경우, 검토해야 할 문헌과 문헌사이의 유사치는 $\binom{n}{2} = \frac{1}{2}n(n-1)$이며, 〈圖 4〉의 유사행렬에서 역삼각형으로 배열되어 있다.

이 행렬을 이용하여 클러스터링을 구성하기 위해서는 다음과 같은 간단한 알고리즘이 필요하다.

① 각 클러스터가 하나의 문헌만을 갖고 있는 n개의 클러스터들로 시작한다.

② 문헌-문헌 유사행렬에서 가장 큰 유사치를 발견한다.

③ 클러스터의 수를 하나 줄이고, 가장 큰 유사치와 관계가 있는 行과 列을 합쳐, 축소된 문헌-문헌 유사행열을 만든다.

④ ②와 ③을 (n-1)만큼 되풀이 한다(이 시점에서 모든 문헌은 하나의 클러스터로 형성된다).

〈圖 4〉에서 가장 큰 유사치는 .83으로서, D4와 D7의 類似關係를 말하고 있다.

다시말해서, 전체문헌 중 D4와 D7이 가장 유사한 주제를 갖고 있기 때문에 이 두 문헌이 가장 먼저 합쳐지는 것이다. 이 두 문헌의 유사치와 다른 문헌들

〈圖 5〉　　문헌-문헌유사행렬의 축소과정 ( )속의 숫자는 가중치.

②	D1	D2	D3	D4/D7	D5	D6	D8	D9
D10	.67	.00	.17	.17	.67	.43	.58	.50
D 9	.29	.14	.14	.36	.50	.50	.58	
D 8	.57	.25	.29	.28	.38	.58		
D 6	.25	.29	.14	.32	.25			
D 5	.43	.13	.33	.32				
(2) D4/D7	.32	.74	.42					
D 3	.33	.40						
(1) D 2	.13							

③	(1) D1	D3	(3) D4,D7 D2	D5	D6	D8	D9
D10	.67	.17	.11	.67	.43	.58	.50
D 9	.29	.14	.29	.50	.50	.58	
D 8	.57	.29	.27	.38	.58		
D 6	.25	.14	.31	.25			
D 5	.43	.33	.25				
(3) D4,D7 D2	.25	.41					
D 3	.33						

④	(2) D1, D10	D3	(3) D4,D7 D2	D5	D6	(1) D8
(1) D9	.40	.14	.29	.50	.50	.58
D8	.58	.29	.27	.38	.58	
D6	.34	.14	.31	.25		
D5	.55	.33	.25			
(3) D4,D7 D2	.18	.41				
D3	.25					

⑤	(2) D1, 10	D3	(3) D4, 7,2	D5	D6
(2) D8,9	.49	.22	.28	.44	.54
D6	.34	.14	.31	.25	
(1) D5	.55	.33	.25		
(3) D4,7,2	.18	.41			
D3	.25				

⑥	(3) D1,10 5	(1) D3	(3) D4,7, 2	(1) D6
(3) D 8,9	.47	.22	.28	.54
(1) D 6	.31	.14	.31	
(3) D4,7,2	.20	.41		
(1) D 3	.28			

⑦	(3) D1,10 5	(1) D3	(3) D4,7, 2
(3) D8,9,6	.42	.12	.287
(3) D4,7,2	.20	.41	
D3	.28		

⑧	(6) D1,10,5 8, 9,6	(3) D4,7,2
(1) D 3	.23	.41
(3) D4,7,2	.25	

⑨	D1,10,5 8, 9,6
D4,7,2,3	.24

과의 유사치를 평균하여 합쳐서 〈圖 5〉의 ②와 같이 첫번째로 축소된 문헌-문헌유사행렬을 얻을 수 있다.

〈圖 4〉에서 D1과 D4, D7의 유사치는 .25와 .38이다. 위에서 말한대로 25와 38을 합하여 2로 나눈 數, 즉 .32가 D1과 (D4, D7) 클러스터와의 유사치로서, 이런 과정을 되풀이하여 〈圖 5〉의 ②가 만들어졌다. 다시〈圖 5〉의 ②에서 가장 큰 유사치는 .74로서 D2와 〈D4, D7〉의 유사관계를 말하고 앞의 과정을 되풀이하여 한 단계 더 축소된 문헌-문헌 유사행렬인 〈圖 5〉의 ③을 얻게되며, 전체문헌에 대한 클러스터링이 구성될 때까지 이 과정이 반복되는 것이다(n-1번). 이 과정에서 우리는 클러스터에 속한 문헌수와 같은 숫자를 그 클러스터의 가중치로 이용하였다. 이 클러스터링 과정의 결과는〈圖 6〉과 같이 "dendrogram"이라 불리우는 系圖형태로 나타낼 수가 있으며, 〈表 2〉는 클러스터 구성 과정을 자세히 보여주고 있다.

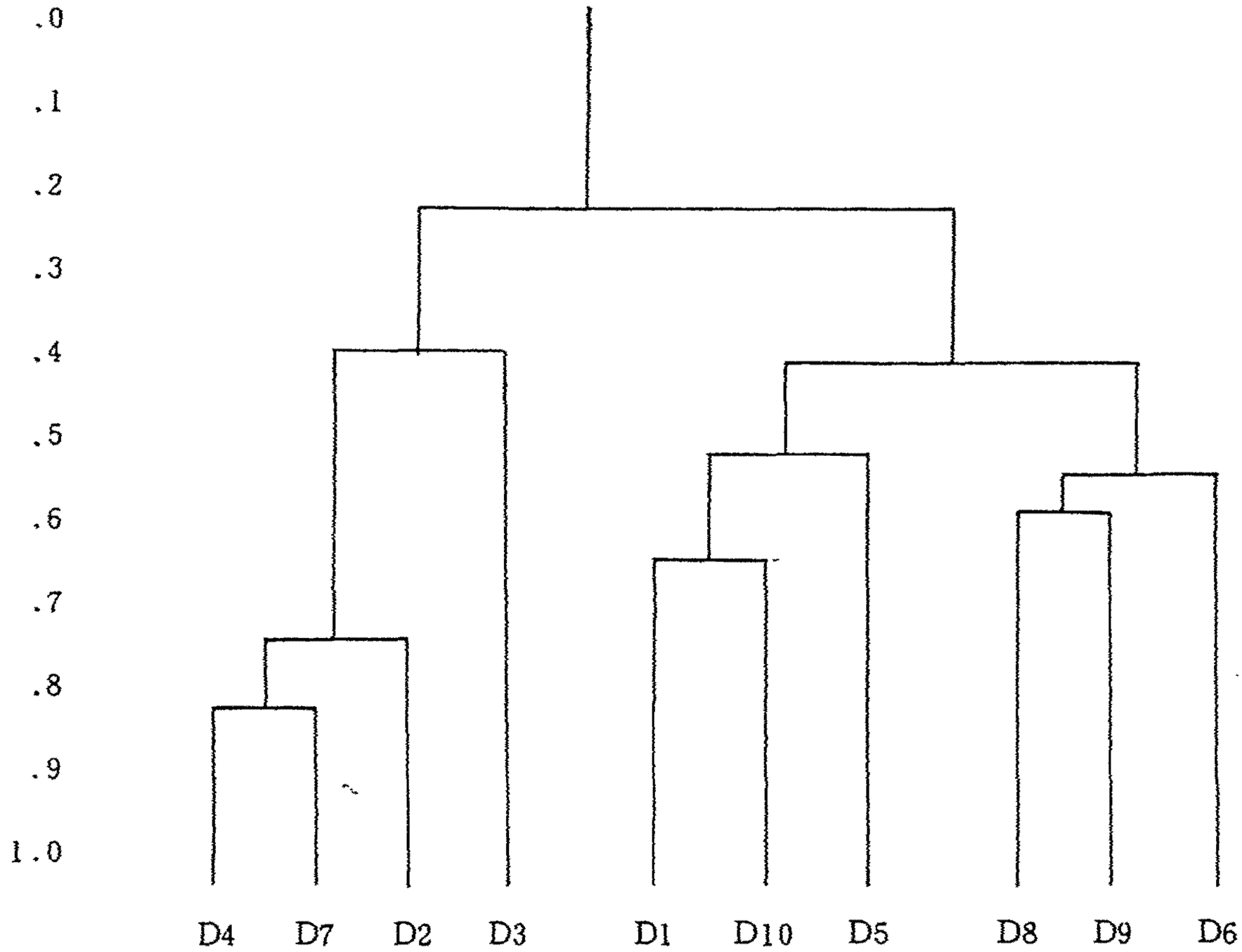
여기서, 예를 든 클러스터링 方法은 수많은 클러스터링方法 중 아주 간단한 예에 지나지 않는다.

우리는 이 例에서 두 그룹이 합쳐지기 위하여 가중치를 이용한 산술평균방법을 취했지만, 가중치를 이용하지 않을 수도 있고, 또는 가중치를 이용한 기하평균방법을 이용할 수도 있다. 나아가 여기에서는 가장 큰 유사치를 이용하였지만, 작은 유사치를 이용할 수도 있고, 더 복잡한 방법도 적용시킬 수 있는 것이다. 일반적으로 클러스터의 구성 및 클러스터 사이의 관계는 두 그룹이 어떻게 합쳐지느냐에 따라 어느 정도 변할 수 있다.

〈表 2〉 〈圖 5〉의 클러스터 형성과정

Cluster 수	Cluster
10	(D1), (D2), (D3), (D4), (D5), (D6), (D7), (D8), (D9), (D10)
9	(D4,D7), (D1), (D2), (D3), (D5), (D6), (D8), (D9), (D10)
8	(D4, D7, D2), (D1, D10), (D3), (D5), (D6), (D8), (D9)
7	(D4, D7, D2), (D1, D10), (D3), (D5), (D6), (D8), (D9)
6	(D4, D7, D2), (D1, D10), (D8, D9), (D3), (D5), (D6)
5	(D4, D7, D2), (D1, D10, D5), (D8, D9), (D3), (D6)
4	(D4, D7, D2), (D1, D10, D5), (D8, D9, D6), (D3)
3	(D4, D7, D2), (D1, D10, D5, D8, D9, D6), (D3)
2	(D4, D7, D2, D3), (D1, D10, D5, D8, D9, D6)
1	(D4, D7, D2, D3, D1, D10, D5, D8, D9, D6)

〈圖 6〉 〈圖 5〉의 dendrogram



Ⅵ. 結 論

이 글에서 우리는 클러스터分析의 일반적인 方法과 이 方法이 情報檢索에 어떻게 應用될 수 있는가를 檢討해 보았다. 이 方法은 情報科學分野에서 광범위하게 利用되고 있는 應用統計學의 한 分野라고 할 수 있다.

우리가 이 글에서 취급하지 못한 클러스터方法들, 특히 그래프理論을 이용한 方法과 集合理論을 이용한 方法은 앞으로 많은 硏究가 되어야 할 것이다. 이 方法들을 文獻檢索시스템에 應用하기 위한 硏究는 最近에 와서 情報科學分野의 硏究者들에게 많은 주목을 받고 있다.

사실, 우리는 아직도 확고한 情報檢索理論이 없다는 것을 느껴왔다. 그래서 많은 硏究者들이 情報檢索理論을 확립하기 위해 確率論에 근거를 둔 檢索理論,

熱力學에서 이용하는 엔트로피(Entropy)를 검색이론으로 도입하는 硏究, 클러스터分析을 통해 시스템 효율을 개선하려는 硏究 등 多角的으로 노력을 기울이고 있다.

이 글이 클러스터分析이 무엇인가를 잘 알지 못하는 情報科學分野의 硏究者들이 클러스터의 槪念을 理解하는데 도움이 되기를 바란다. 아울러, 가까운 장래에 종래의 클러스터링方法보다 더 效率的인 利用者의 질문식에 기초를 둔 클러스터링을 구성하는 方法이 개발되어야 할 것이다.

〈參 考 文 獻〉

1. Anderberg, M. R., *Cluster Analysis for Application*. New York : Academic Press, 1973.
2. Borko, H., and Muller, Charles W., "Automatic Document Classification Ⅱ-Additional Experiments," *JACM*, 1963, 10, pp.151-162.
3. Dillon, H. "The Use of Clustering Technique in the Analysis of Judical Mose," *Information Science*, 1975, 8, pp.95-107.
4. Gordon, A. O. *Classification*, London : Chapman and Hall, 1981.
5. Gotlieb, C. C., and Kumar, S., "Semantic Clustering of Index Terms," *JACM*, vol.15, no.4, 1968, pp.493-513.
6. Johnson, S. C. "Hierarchical Clustering Schemes", *Psychometrika*, vol.32, no.4, 1967, pp.241-254.
7. Kessler, M. H. "Bibliographic Coupling between Scientific Paper", *American Documentation*, vol.14, no.1, 1963, pp.10-25.
8. Kim, Jae-On and Mueller, Charles, *Introduction to Factor Analysis-What It Is and How to Do It*, London : Sage University Press, 1978.
9. Mezzich, Juan E., and Solomon, Herbert, *Taxonomy and Behavioral Science*, New York : Academic Press, 1980.
10. Miyamoto, S., and Nakayama, K., "A Technique of Two-Stage Clustering Applied to Environmental and Civil Engineering and Related Methods of Citation Analysis", *JASIS*, vol.34, no.3, 1983, pp.192-201.
11. Salton, G. *Automatic Information Organization and Retrieval*, New York : McGraw-Hill, 1968.
12. Sokal, R. R., and Rohlf, F. J., "The Comparison of Dendrogram by Objective Methods", *Taxon*, 1962. 11, pp.33-40.

13. Sparck Jones, K., *Automatic Keyword Classification for Information Retrieval*, London : Butterworths, 1971.
14. Van Rijsbergen, C. J., "An Algorithm for Information Structuring and Retrieval", *The Computer Journal*, vol.14, 1971, pp.407-412.
15. _____, *Information Retrieval*, London : Butterworths, 1979.
16. Ward, Jr.J. H., "Hierarchical Grouping to Optimise an Objective Function", *Journal of American Statistics Association*, vol.58, no.301, 1963, pp. 236-244.
17. Yu, Clement Y. "A Clustering Algorithm Based on User Queries, *JASIS*, vol.25, no.4, 1974, pp.216-218.