용어의 자동분류에 관한 연구

A Study on Automatic Keyword Classification

서 은 경

초 록

본 논문은 기계가독형 데이타베이스를 탐색하는 자연어 정보검색 시스템에서 검색용 디소오러스를 이용하면 정보검색효율이 향상된다는 전제하에, 검색용 디소오러스 자동 작성방법 중의 하나인 용어 자동분류를 우리말 용어에 적용시켜 실험하였고, 이 결과로 형성된 용어군의 응용방법을 제시하였다.

용어자동분류는 용어사이에 존재하는 어의적 관계가 한 문장에서 출현하는 용어의 통계적 양상에 근거하여 밝혀질 수 있다는 가설하에 세워진 방법으로, 본 논문에서는 심리학 분야의 국내 학술잡지중 초록이 수록된 4개의 잡지를 실험대상잡지로 선택하였다.

Abstract

In this paper, the automatic keyword classification which is one of the automatic construction methods of retrieval thesaurus is experimented to the Korean language on the basis that the use of retrieval thesaurus would increase the efficiency of information retrieval in the natural language retrieval system searching machine-readable data base. Furthermore, this paper proposes the application methods.

In this experiment, the automatic keyword classification was based on the assumption that semantic relationships between terms can be found out by the statistical patterns of terms occurring in a text.

Ⅰ. 서 론

학문과 기술이 다각적으로 발전함에 따른 정보량의 급격한 증가와 특정적이며 복합적인 이용자의 정보요구로 인하여 문헌자료와 정보를 체계적으로 수집·정리·보존하여 이용자에게 신속·정확하게 제공하는 정보검색시스템의 필요성이 크게 부각되었으며, 또한 최근에는 전문가가 아닌 이용자가 직접 자연어로 검색할 수 있는 자연어

* 도미 유학중
접수일자 : 1984. 11. 27.

정보검색시스템이 개발되었다.

이로써 이용자는 시스템과 계속적인 상호작용을 하면서 손쉽게 정보를 검색할 수 있게 되었으나, 자연어를 사용하는 정보검색시스템이 여러 성능평가 실험에서 비교적 검색효율이 낮은 것으로 밝혀졌다[1]. 이는 검색어인 자연어가 색인작성자와 이용자 사이에 통일되기 힘든 특정적인 성격을 띨 뿐만 아니라 검색시 동의어·동형이의어·대소개념어·상하관련어 등의 조절이 불가능하기 때문에 나온 결과이다. 따라서 이러한 자연어 검색에서 오는 문제점을 해결하여 정보검색효율을 높혀줄 수 있는 검색용 디소오러스 작성이 필요하게 되었다.

본 논문에서는 자연어 검색시스템에서 검색용 디소오러스를 이용하면 정보검색효율이 향상된다는 전제하에 그 작성 방법중의 하나인 용어자동분류를 우리말에 적용시켜 실험하였고, 이 실험결과를 가지고 아직 미흡한 연구상태인 우리말 디소오러스 작성에 있어서 필요한 기초자료를 제시해 보고자 한다.

Ⅱ. 이론적 배경

1. 용어자동분류의 의의

분류란 여러 특성을 지닌 구성원들의 집합을 주어진 한 특성에 따라 여러개의 부분적인 집단으로 나누는 것이다. 또한 분류과정은 유사성이라는 기준을 가지고 각각의 개체를 나누는 것이므로 분류작업은 사실상 동물의 분류에서와 같이 이미 오래 전부터 시행되어 왔다[2]. 그런데 분류대상이 개념을 나타내는 용어인 용어분류의 경우는 최근에 정보검색 시스템이 부각되면서 정보검색에 없어서는 안될 중요하고 기본적인 작업의 하나가 되었다.

더구나 특정한 개념을 나타내는 전문어 이외에 용어가 과정·활동의 개념까지 포함하는 넓은 범위의 자연어 일때, 용어분류에는 전문가의 높은 지적 수준과 막대한 시간적 투자가 요구된다. 즉 수작업으로 많은 양의 용어를 분류하는 경우 그 소요되는 시간은 막대하며[3], 또한 주제전문가나 분류전문가가 분류를 한다 해도 항상 객관적으로 일관성 있는 분류결과를 갖게 되기는 어려운 일으므로 이를 해결하기 위하여 전문가들은 기계적 처리가 가능한 분류방법에 관심을 기울이게 되었다.

용어의 자동분류는 용어의 문맥적인 성질을 근거로 용어의 의미를 결정하여 동의어·관계어군을 자동적으로 만드는 방법[4]으로 이 작성방법으로는 컴퓨터를 통한 완전자동어의적 분석방법과 통계에 근거하는 방법, 두가지가 있다. 이중 완전자동어의적 분석방법은 사실상 사용되지 않고 있으며, 용어가 한 문헌에 출현하는 통계에 근거하는 방법만이 이용되고 있다. 말하자면 컴퓨터는 용어가 지닌 의미를 직접적으로 인식할 수는 없지만 각각 문장에서 나타나는 용어의 양상을 통계적으로 파악해서 각 용어쌍의 어의적 관계를 발견할 수가 있는 것이다. 즉 용어 a나 b가 각각 용어 c와 함께 매번 출현한다면 용어 a와

1) F.W. Lancaster, et al., "Evaluating the Efficientiveness of On-Line Natural Language Retrieval System," *Information Storage and Retrieval*, Vol. 8, No. 5 (Oct., 1972), pp. 223~245.

2) Robert R. Sokal, "Clustering and Classification: Background and Current Direction," edited by J. Van Ryzin *Classification and Clustering* New York, Aeademic Press, 1977, p. 1.

3) Dagobert Soergel, *Indexing Language and Thesanri: Constnction and Mainteriance*, LA: Melville, 1974, p. 15.

4) M.E. Lesk, "Word-Word Association in Document Retrieval System," *JASIS*, Vol. 20, No. 1, (Jan., 1969), pp. 27~29.

b는 의미적으로 밀접한 관계가 있다는 가설과 용어 p와 q가 항상 같이 출현한다면 그 문헌을 검색할 경우 용어 p대신 q를, q대신 p를 이용하여 검색해도 같은 문헌이 나올 수 있다는 가설[5]을 이용하여 같은 주제에 대한 동의어군과 다른 각도에서 같은 주제를 접근하는 관련어군을 만들 수가 있는 것이다.

이러한 용어자동분류는 용어간의 어의적 특성을 용어가 지닌 의미에서가 아니라 문맥적 속성에서 찾는다는 것과 한 용어가 오직 하나의 집합군에만 속하는 것이 아니라 여러군에 중복되어 비계층적으로 나타날 수 있다는 특징을 지니며, 이의 궁극적인 목적은 검색시스템에서 검색용 다소오러스로 사용하여 하나의 색인어만을 이용하는 경우보다 검색의 폭을 확장시켜 검색힘으로써 검색효율을 높히는데 있다.

2. 용어자동분류의 발전

이용자의 질문과 문헌의 내용을 나타내는 용어가 서로 일치하는 범위를 넓히기 위해 각각의 용어들을 분류시키는 방법인 용어자동분류는 다음과 같은 두단계를 거쳐서 이루어진다. 첫번째 단계는 용어간의 동시출현 빈도수에 따른 유사성의 측정으로 유사도 측정공식을 이용하여 용어간의 유사관계를 수치값으로 나타내는 과정이다. 두번째 단계는 측정된 유사계수 값을 이용하여 연관성이 높은 용어를 모아주는 알고리즘을 적용시켜 같은 집합군에 유사한 용어가 모아지게 하는 과정을 말하는데 이때 형성된 집합군을 유사어군으로 간주한다[7].

이에 따라 용어자동분류에 시도된 방법도 유사성 측정방법과 유사성 측정후 용어군을 형성하는 알고리즘에 관한 연구로 나뉘어져서 발전되어 왔다.

2.1 유사성 측정 방법

용어의 자동분류는 두 용어간의 상호관계에 기초하여 행해지는데 이 관계를 나타내는 척도를 유사성, 연관성, 비유사성 등으로 표현[8]하고 있으며, 두 용어간의 유사성을 측정하여 수치로 표현한 것을 유사계수라 한다. 이 유사계수는 두개의 대상물이 한 집합의 일원인 경우, 그들의 관계를 측정한 값으로 0에서 1까지의 범위를 가지고 있다[9].

1950년 말, 볼(Ball)의 유사성 측정방법 연구를 시초로 측정방법이 여러 연구자들에 의해 연구되기 시작하여 타니모토(Tanimoto)에 의하여 최초로 실제 이용될 수 있는 함수식이 나왔다. 그후 유사계수 측정방식이 계속 제시되었으며 이중 실제 많이 이용되고 있는 유사성 측정공식은 타니모토 유사도 측정공식, 논리적 중복계수, 코싸인 상관계수 등이다.

타니모토의 공식은 대상물 x, y간의 유사계수를 집합원리를 이용하여 측정한 것으로 다음과 같다[10].

$$S(x,y)=\frac{x\cdot y \text{ 동시 출현수}}{(x\text{의 출현수}+y\text{의 출현수})-(x\cdot y \text{ 동시 출현수})}$$

$$=\left|\frac{\phi(x)\cap\phi(y)}{\phi(x)\cup\phi(y)}\right|$$

5) Karen Spark Jones and Martin Key, *Linguistic and Information Science*, NY: Academic Press, 1973, pp. 162~163.
6) Karen Sparck Jones, *Automatic...*, pp. 16~25.
7) M. Dillon and D. Caplan, "A Technique for Evaluating Automatic Term Clustering." *JASIS*, Vol. 31, No. 2 (March, 1980), p. 89.
8) C.J. Van Rijsbergen, *Information Retrieval 2nd ed.*, London: Butterworths, 1979, p. 38.
9) Brian Everitt, *Cluster Analysis*, NY: John Wiley & Sons, 1974, pp. 50~51.
10) D. Roger and T. Tanimoto, "A Computer Progran for Classifying Plants," *Science*, Vol. 132, No. 3432(Oct., 1962), pp. 1115~1118.

(이때 $|\phi|$는 ϕ군의 구성원수를 가르친다)

이 공식은 스파크존스(Sparck Jones), 오거스톤(Auguston)[11], 고틀리브(Gotlieb)에 의해 이용되었으며, 특히 고틀리브와 쿠머(Kumar)는 이 공식을 이용하여 두 용어 $x \cdot y$간의 거리를 나타내는 비유사도 측정공식을 만들었다[12].

또한 논리적 중복계수(Logical Overlap Coefficient: LO)와 코싸인 상관계수(Cosine Correlation Coefficient: CC)는 SMART(System for the Mechanical Analysis and Retrieval of Text) 프로젝트에서 사용된 공식으로 다음과 같다[13].

$$LO=\frac{\sum_{i=1}^{t} \min(v_i \cdot w_i)}{\min(\sum_{i=1}^{t} v_i \cdot \sum_{i=1}^{t} w_i)}$$

$$CC=\frac{\sum_{i=1}^{t} v_i\, w_i}{\left[\sum_{i=1}^{t} (v_i)^2 \cdot \sum_{i=1}^{t} (w_i)^2\right]^{1/2}}$$

이 방식은 용어를 하나의 벡터로 나타내는 방법으로 v, w는 t-디멘션 벡터를 말한다.

유사계수에 대한 또하나의 실험으로는 중요하다고 생각하는 용어에 가중치를 줌으로써 더욱 유사한 용어들이 모이도록 한 방식이 있다. 용어에 가중치를 주는 방법으로는 빈도수를 이용하여 가중치를 주는 방법(Statistical Weighting Scheme)과 분류의 목적으로 분류자가 중요하다고 생각하는 용어에 상대적인 가중치를 주는방법[14),15)]으로 나눌 수 있는데 가중치 값을 매기는 척도의 일관성 문제 및 그 타당성 여부로 여러 연구자들 사이에 논란이 많으므로 많이 이용되고 있지 않다[16].

2.2 용어의 분류방법

기계적 처리로 디소오러스를 작성하려는 시도는 1958년에서 1966년 사이에 데이타베이스 시스템이 만들어짐에 따라 정보검색시 검색용 디소오러스의 자동 작성방법에 대한 연구로써 시작되었다. 특히 용어의 통계적 속성으로 용어의 유사성을 측정하여 기계적 처리로 디소오러스를 구성하는 방법은 매우 많은 연구자들에 수행되었는데 대표적인 연구자들로는 베이커(Baker), 보코(Borko)[17], 도일(Doyle)[18], 길리아노(Giuliano)[19], 스파크존스, 마론과 쿤(Maron and Kuhns), 니드햄(Needham)[20], 스타일(Stiles)[21] 등을 들 수 있다.

본 장에서는 가장 많은 연구가 수행된 그래프 이론과 클럼프 이론을 중심으로 실제 용어분류를 형성하는 알고리즘들을 개관해 보기로 한다.

2.2.1 그래프 이론 방법

그래프 이론방법이란 용어간의 유사성을 측정

---

11) J.G. Auguston and J. Minker, "Derving Term Relationship for a Corpus by Graphic Theoretical Cluster," *JASIS*, Vol. 21, No. 2(March, 1970), p. 101.

12) C.C. Goflieb and S. Kumar, "Semantic Clustering of Indexing Terms," *J. of ACM*, Vol. 15, No. 4 (Oct., 1968), p. 495.

13) G. Salton ed., *The SMART Retrieval System*, NJ: Prentice Hall, 1971, pp. 163~166.

14) G. Salton and M.I. McGrill, *Introduction to Modern Information Retrieval*, N.Y.: McGraw-Hill, 1983, pp. 132~133.

15) 샐톤은 이 두가지 방법을 $\text{Weight}_{ik}=\text{FREQ}_i/\text{DOCFREQ}_k$ $\text{Weight}_k=\text{FREQ}_{ik}\cdot\text{DESCVALUE}_{ik}$로 표시하고 있다.

16) Brain Everitt, pp. 49~50.

17) H. Borko and M. Bernick, "Automatic Document Classification," *J. of ACM*, Vol. 10 and 11, No. 2 (April, 1963, 1964), pp. 151~162; 138~151.

18) L.B. Doyle, "Indexing and Abstracting by Associative," *American Documentation*, Vol. 13, No. 4 (Oct., 1962), pp. 378~390.

19) V.E. Giuliano and P.E. Jones, "Linear Associative Information Retrieval," *Information Handling*, Vol. 1 (1963).

20) R.M. Needham, "Application of the Theory of Clumps," *Mech. Transl. Comp. Linquist*, Vol. 8 (1965), pp. 113~127.

21) H.E. Stiles, "The Association Factor in Information Retrieval," *J. of ACM*, Vol. 8, No. 2 (April, 1961), pp. 271~279.

할 수 있는 매칭함수를 이용하여 그래프로 용어군을 만드는 것을 말한다. 즉 매칭함수에 의해서 계산된 유사계수가 두 용어간의 유사성 정도를 나타낸다고 보고, 유사계수에 일정한 기준치를 정해준 다음 그 기준치 이상의 용어를 연결시켜 용어군을 형성하는 방법이다[22].

이와같이 그래프형태로 나타난 용어군을 그 형태별로 나누어 보면, 용어군이 시작하는 대상물에서 계속 일련선상으로 이어지는 형태인 스트링(String)과 각 대상물이 용어군에 속해있는 다른 대상물과 최소한 한번 이상 연결되어 그 집합군의 속성을 최대로 나타내는 부분연결 그래프(Connected Component), 각 대상물이 그 집합군에 있는 모든 대상물과 연결된 상태를 말하는 완전연결 그래프(Maximal Complete Subgraph)로 나타낼 수 있다[23].

그래프 이론을 이용하여 용어군을 발견하는 알고리즘은 1950년 말경 쿤에 의하여 처음으로 사용되었다[24]. 그러나 그의 알고리즘은 컴퓨터에 이용할 수 있는 성질의 것이 아니었고 그후 갤러(Galler)와 휘셔(Fisher)가 부분연결 그래프 방법을 적용시켜 자동적으로 용어군을 형성시켰으나 용어군에서 용어간의 연관성이 잘 나타나지는 못했다[25].

해리(Harry)와 로스(Ross)[26]는 처음으로 쿤의 알고리즘을 변형시켜 컴퓨터에 적용시켰고 비얼스톤(Biersfone)[27]는 완벽한 하부그래프(Complete Subgraph)를 이용하여 완전연결 그래프를 만드는 알고리즘을 개발하였다.

보너(Bonner)[28]는 규모가 큰 데이타에 적합한 알고리즘을 연구하여 그 방법을 개발하였다. 즉 먼처음 한 대상물에 대한 완전연결 그래프를 형성한 후, 기준치 i 값을 정하여 대상물이 들어올 적마다 연결여부를 결정하여 매번 집합군을 새로 형성하는 방법이다.

고틀리브와 쿠머는[29] 완전연결 그래프로 형성된 용어군들이 서로 너무 밀접하게 연결된 상태를[30] 완화시키며, 부분연결 그래프 보다는 용어군에 좀 더 유사한 용어가 모이도록 하는 방법을 연구하였다. 즉 그는 완전연결 그래프로 용어군을 형성한 다음 이를 다시 재형성하여 앞서 말한 상태의 용어군을 형성하는 알고리즘을 개발하였다. 그들은 재형성된 용어군을 확대개념용어군(Diffuse Concept)라 하였고 이 용어군은 완전연결 그래프로 형성된 용어군간의 거리를 측정하여 그 거리값이 일정 기준치 보다 작거나 같을 때 이 두개의 용어군을 결합하여 형성되며, 용어군간의 거리측정은 다음과 같은 공식을 사용하였다.

$$d_{ij}=1-|C_i\cap C_j|/|C_i\cup C_j|^*$$

22) C.J. Van Rijsbergen, p. 48.
23) Ibid., pp. 48~49.
24) J.G. Auguston and J. Minker, p. 102.
25) J.G. Auguston and J. Minker, "An Analysis of Some Graph Theoretical Cluster Technique," *J. of ACM*, Vol. 17, No. 4 (Oct., 1974), p. 576.
26) J.G. Auguston and J. Minker, p. 102.
27) J.G. Auguston and J. Minker, "An Analysis ...." p. 576.
28) R.E. Bonner, "On Some Clustering Technique," *IBM Journal*, Vol. 8, No. 1 (Jan., 1964), pp. 22~32.
29) C.C. Gotlieb and S. Kumar, "Semantic...." p. 496.
30) 이 예로 오거스톤과 민커의 실험을 들 수 있다. 그들은 비얼스톤 알고리즘을 이용하여 용어자동분류한 결과 8개의 분류군으로 나뉘어 졌는데 그중 64개 용어가 포함된 분류군이 세개이며, 이 세개군에 공통적으로 나타난 용어의 수는 63개이었다.

* C_i, C_j는 핵심개념용어군, d_{ij}는 두 용어군(C_i와 C_j)의 거리
$|C_i\cup C_j|$는 두 핵심개념용어군에 포함된 모든 용어의 수
$|C_i\cap C_j|$는 두 핵심개념용어군에 공통적으로 포함된 용어의 수

한편 스파크 존스[31]는 위의 연구와는 달리 집합군을 스타(Star), 스트링, 클리끄(Clique), 클럼프(Clump) 4가지로 정의내리고 있는데 실제 스타형태는 일반적으로 규모가 큰 데이타 집단형태라 할 수 있으며 클리끄는 앞서 말한 완전연결 그래프로 일컬어지므로 그의 연구에 대해서는 클럼프를 중심으로 다음 장에서 살펴보기로 한다.

2.2.2 클럼프 방법

클럼프 방법은 한 용어의 속성이 여러개로 나타날 때 이 용어가 다른 용어군에도 중복되어 나타날 수 있게하는 분류방법이다. 클럼프라는 용어는 1960년 존(John)과 니드햄[32]에 의해서 처음으로 소개되었으며 클럼프 이론은 용어군 안에 있는 용어가 용어군 밖에 있는 용어보다 용어사이의 연관성이 더 크다는 가설아래, 이미 형성된 클럼프와 그 클럼프에 속하지 못한 잔여 집합사이와의 경계범위, 클럼프와 그 나지지 대상물과의 분리된 정도에 기초한 분류방법이다. 이때 클럼프와 그 잔여 집합과의 관계는 응집함수(Cohension Function)에 의하여 표시되는데 이 응집함수는 두 집합간의 연결강도(Strength of Connection)를 수치로 나타낸 것이다[33].

클럼프는 응집함수에 의해 각각의 용어를 두개의 그룹으로 분산시킴으로서 형성되며, 이때 용어의 분산은 두 집합군 사이의 응집함수를 연속적으로 최소화 시킴으로써 이루어진다[34]. 클럼프를 형성하기 위한 응집함수의 실험은 많은 연구자[35]들에 의하여 연구되었다.

Ⅲ. 우리말 용어의 자동분류 실험

1. 실험집단 구성

본 실험에서는 심리학 분야의 학술잡지 중에서 초록을 포함하는 「한국심리학회지」 「임상심리학보」 「심리학 연구」 「사회심리학 연구」를 실험대상잡지로 선택하였다. 심리학의 여러 하부주제 중 사회심리학, 임상심리학, 발달심리학과 관련된 논문기사 중에서 특히 연구·개발·실험 평가의 성격을 띈 34편의 기사를 선정하여 그 초록을 구성하는 용어들을 자동분류하였다.

34개로 초록의 수를 한정한 이유는 행렬의 크기를 너무 방대해지지 않도록 하기 위해서이며 선정된 초록의 길이는 평균 479자이고, 입력된 초록의 총길이는 166,277자이다.

실험집단의 구성은 이미 선정된 초록을 컴퓨터에 입력시켜 실제 실험에 이용되는 용어들을 추출하는 과정이다. 초록을 컴퓨터에 입력시키기 전에 먼저 용어특성을 살펴본 결과 특정성이 매우 높을 뿐만 아니라 같은 개념이 여러 용어로 표현된 것이 많아 용어의 동시출현 빈도수가 매우 낮았다.

본 실험에서는 이를 해결하고 작은 실험집단의 규모로 뚜렷한 유사어군을 형성하기 위하여 초록에 수록된 용어들을 통제하여 선정하였는데, 그 방법은 다음과 같다.

첫째, 입력된 초록중 국한문혼용체인 것은 한문을 한글로 변형시켜 입력하였다.

둘째, 초록을 구성하는 용어 중에서 기능어(조사, 전치사, 접속사)와 상용어(예 : 연구, 고찰…), 그리고 주제와 관련없는 형용사 및 동사(예; 높은, 본, 나타나다…) 등을 제외한 모든 용어를 선정대상으로 하였다.

31) Karen Sparck Jones, *Automatic...*, pp. 55~63.
32) R.M. Needham, Application..., pp. 113~127.
33) K. Spark Jones, "Clump, Theory of," *Encyclopedia of Library and Information Science*, Vol. 5, N.Y.: Marcel Dekker, 1969, pp. 208~224.
34) Brain Everitt, *Cluster...*, p. 36.
35) 스파크존스, 니드햄, 파커(Parker), 로도(Rhodes), 잭슨(Jackson)

세째, —적, —성, —형으로 끝나는 형용사는 불용어로 보지않고 수식받는 명사와 함께 한개의 용어로 처리하였다(예 : 물질적 보상, 역행성 기억상실, Y형미근), 단 이와 같이 명사를 수식하는 형용사가 주제를 나타내는지를 알아보기 위해서 전체 10%만을 분리하여 각각 한 용어로 간주하였다.

네째, 같은 의미의 복합어인데 표기하는 방식이 다른 경우에는 한가지 방식으로 통일시켰다. (예; Y-미근, Y형미근⇒Y형미근)

다섯째, 복합어 표기시 띄어쓰기를 다르게 한 경우 띄어쓰기를 조절하여 같은 용어로 추출되도록 하였다. (예; 화친차원 인상, 화친차원인상⇒화친차원인상)

여섯째, 외래어가 우리말을 수식해 줄 때 이 두 단어는 붙여쓰고 한 용어로 처리하였다. (예; P/O관계, U형곡선) 또한 외래어의 복합어도 한 용어로 처리하였다. (예; Test of Irrelevance)

일곱번째, 초록에 나타난 외래어는 원래대로 표기되거나 발음되는대로 또는 번역하여 쓰여졌는데 이 중 같은 용어가 다른 형태로 쓰여진 경우 선정된 초록에서 가장 많이 쓰여진 형태로 조정하였다, (예; Rorcharch, 로샤검사⇒로샤검사)

이와 같은 방법에 의하여 선정된 용어에 특정기호[36]를 표시하여 입력시켰다. 그 결과 추출된 용어는 1,159개이었고, 이중 중복된 단어를 제외시켜 총 320개의 용어를 실험집단으로 구성하였다.

2. 실험방법

2.1 용어간의 유사성 측정

본 실험의 첫번째 단계는 각 초록에서 추출된 용어간의 유사성을 측정하는 일이다. 두 용어간의 유사도는 용어의 빈도수와 동시출현 빈도수의 비율로 나타나는 유사계수 값을 이용하여 측정할 수 있다.

용어간의 유사성을 측정하는 공식은 자연어 구분일 경우 타니모토유사도 측정방법이 효율적이라는 딜론(Dillon)과 캡프랜(Caplan)[37]의 실험결과를 근거로 하여 다음과 같은 식을 이용하였다.

$$S(x,y)=\frac{x\cdot y \text{ 동시 출현수}}{(x\text{의 출현수}+y\text{의 출현수})-(x\cdot y \text{ 동시 출현수})} \quad \cdots\cdots(1)$$

위 식의 결과물을 N이라 할 때 N은 $n\times n$ 대칭행렬로 형성되며 $S(x,y)$의 값은 0에서 1사이의 범위를 지니게 된다.

2.2 부분연결 그래프 이론을 이용한 용어군 형성

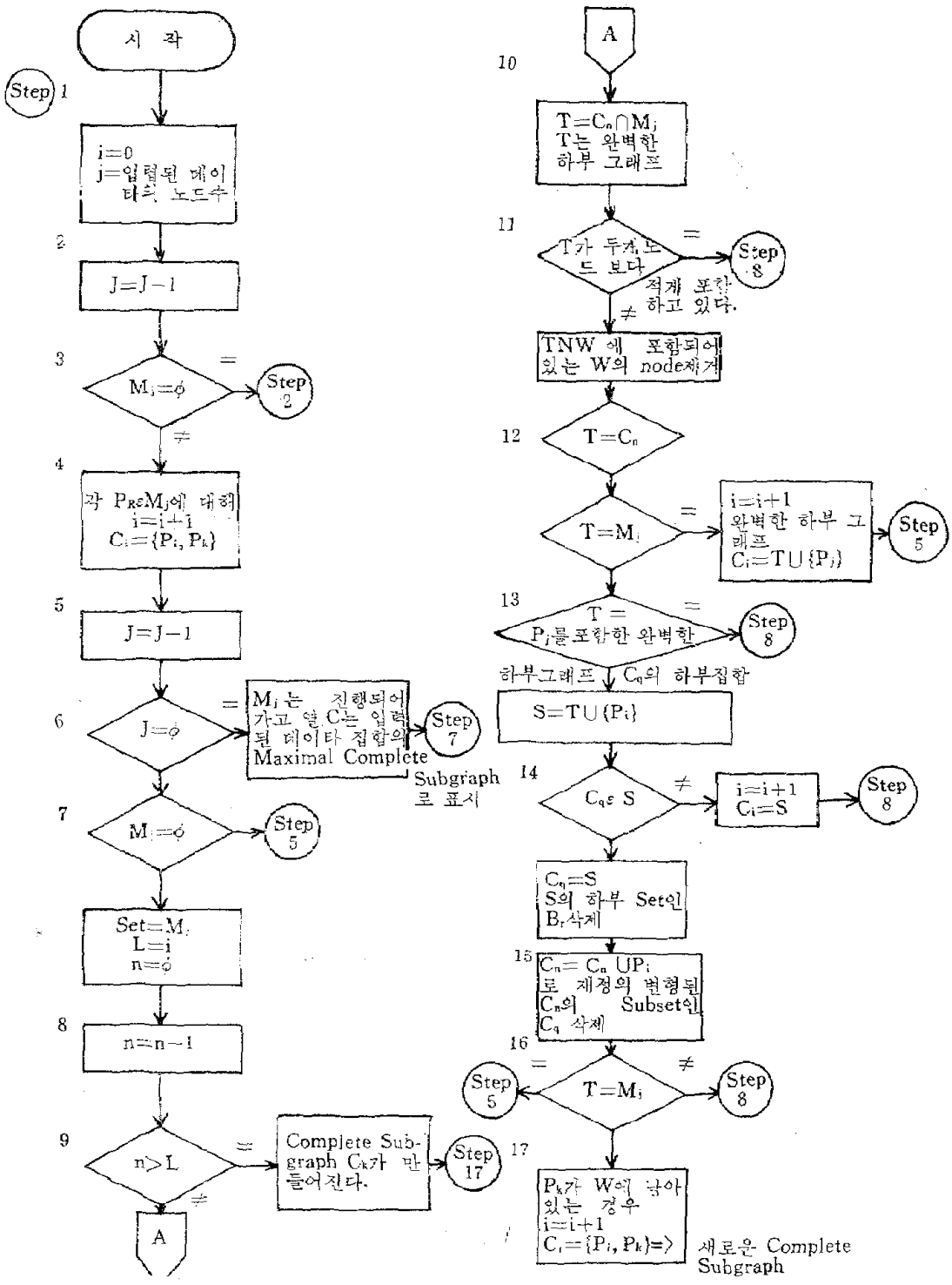
본 실험에서는 부분연결 그래프 방법을 적용시켜 수작업으로 용어군을 형성하였으며, 그 형성과정은 다음과 같다.

먼저 용어군의 규모를 정해주는 기준치를 정한 다음, 임의의 한 용어를 기준으로 하여 이 용어와 기준치 값보다 같거나 큰 유사계수를 가진 용어들로 일차적인 용어군을 형성한다. 다시 용어군에 속해있는 각각의 용어와의 유사계수 값이 기준치 값보다 같거나 큰 용어들을 연쇄적으로 이미 형성된 용어군에 포함시켜 더 이상 이 용어군에 용어가 포함되지 않을 때까지 계속하여 한 개의 용어군을 형성한다. 다음 이용어군에 속해있지 않는 다른 한 용어를 기준으로 삼고 위의 과정을 반복하여 또 다른 용어군을 형성한다. 이 반복되는 과정은 모든 용어가 각각 용어군에 모두 포함될 때에 끝나게 된다.

이런 과정을 통하여 형성된 용어군은 다음과 같은 성격을 띄게된다. 용어군에 속해있는 모든

36) 본 실험에서는 백슬라쉬(back slash)를 이용하였다.

37) M. Dillon and D. Caplan, "A Technique...," p. 89.


시 작
Step 1
i=0
j=입력된 데이타 노드수
2
J=J-1
3
$M_j=\phi$
=
Step 2
≠
4
각 $P_k \varepsilon M_j$에 대해
i=i+1
$C_i=\{P_i, P_k\}$
5
J=J-1
6
J=φ
=
M_j는 진행되어 가고 열 C_i는 입력된 데이타 집합의 Maximal Complete Subgraph로 표시
Step 7
7
$M_j=\phi$
Step 5
Set=M_j
L=i
n=φ
8
n=n-1
9
n>L
=
Complete Subgraph C_k가 만들어진다.
Step 17
≠
A
A
10
$T=C_n \cap M_j$
T는 완벽한 하부 그래프
11
T가 두개 노드 보다 적게 포함하고 있다.
=
Step 8
≠
TNW에 포함되어 있는 W의 node제거
12
$T=C_n$
T=M_j
=
i=i+1
완벽한 하부 그래프
$C_i=T \cup \{P_j\}$
Step 5
13
T=P_j를포함한 완벽한 하부그래프
=
Step 8
C_q의 하부집합
S=T∪{P_i}
14
$C_q \varepsilon S$
≠
i=i+1
C_i=S
Step 8
C_q=S
S의 하부 Set인 B_r삭제
15
$C_n=C_n \cup P_i$
로 재정의 변형된 C_n의 Subset인 C_q 삭제
16
T=M_j
=
Step 5
≠
Step 8
17
P_k가 W에 남아 있는 경우
i=i+1
$C_i=\{P_i, P_k\}$=>
새로운 Complete Subgraph

용어는 최소한 그 용어군에 속해있는 한 개 이상의 다른 용어와의 유사계수 값이 기준치보다는 같거나 크며, 그 용어군에 속해있지 않는 용어들과의 유사계수 값은 모두 기준치보다 작다는 성격을 지니게 된다. 이로써 용어군의 속성은 최대로 나타나나 어의적으로 거의 연관성이 없는 용어들이 한 용어군에 포함되는 결점을 지니게 된다.

그러므로 본 실험에서는 실제 디소오러스로 이용될 수 있는 확대개념용어군을 형성하는데 드는 시간을 줄이는 방법으로 이 용어군을 이용하였다. 즉 부분연결 그래프 방식으로 형성된 각각의 용어군 안에서 확대개념용어군을 형성하므로서 서로 연관성이 없는 용어의 비교를 없앨 수가 있었다.

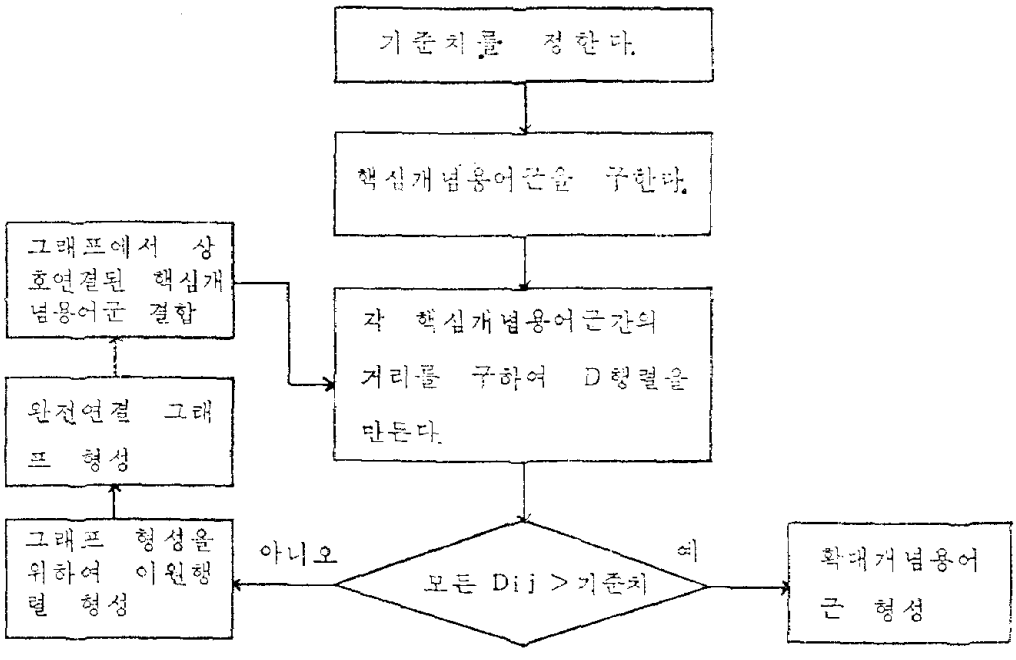
2.3 핵심개념용어군 형성

핵심개념용어군은 완전연결 그래프 방식으로 형성된 용어군으로 이 용어군 안에 있는 모든 용어들과의 유사계수 값은 기준치 값보다 같거나 크며, 유사계수 값이 기준치보다 작은 용어는 존재하지 않는 속성[38]을 갖고 있으므로 매우 밀접한 용어만이 핵심개념용어군에 포함되어 각각의 용어군은 뚜렷한 개념을 나타내나, 많은 용어들이 핵심개념용어군에 중복되어 속해있기 때문에 단일 개념을 지니지 못하는 성격을 지닌다. 그러므로 핵심개념용어군도 직접 디소오러스로 이용되기에는 적합하지 못하고 확대개념용어군을 형성하는데 이용되어진다.

이 용어군의 형성은 보너의 알고리즘[39]을 이용하였으며 그 형성과정은 다음의 흐름도로 나타내었다.

2.4 확대개념용어군 형성

확대개념용어군은 핵심개념용어군의 중복성을 이용한 고틀리브와 쿠머의 이론에 근거하여 형성하였다. 즉 핵심개념용어군 끼리의 중복성 정도를 각 핵심개념용어군의 거리로 보고, 중복성 정도가 크면 핵심개념용어군간의 거리의 차는 작고 이들이 유사하다는 이론 아래 이미 형성된 핵심개념용어군 간의 거리를 측정하여 기준치 값보다 작은 경우 두 용어군을 결합시키는 방법을 이용



〔그림 2〕 확대개념용어군 형성과정 흐름도

38) C.C. Gotlieb and S. Kumar, "Semantic...," p. 495.
39) R.E. Bonner, "On Some...," pp. 32~32.

하였다[40].

이때 핵심용어군과의 거리는 두개의 핵심개념용어군에 중복되어 포함되어 있는 용어의 비율로 나타내지며, 이 공식은 다음과 같다.

$$d_{ij}=1-\left|\frac{C_i\cap C_j}{C_i\cup C_j}\right| \quad \cdots\cdots(2)$$

위 식에서 $|C_i\cup C_j|$는 핵심개념용어군 C_i와 C_j에 포함된 전체용어의 수이며, $|C_i\cap C_j|$는 두개의 핵심개념용어군 C_i와 C_j에 공통으로 나타난 용어의 수를 말한다.

핵심개념용어군을 결합시켜 확대개념용어군으로 만드는 과정을 간단한 흐름도로 살펴보면 다음과 같다.

3. 실험내용

3.1 용어간의 유사성 측정

본 실험의 첫번째 단계로, 320개의 실험집단 용어가 초록에 출현하는 빈도수와 한 문장에 두 용어가 같이 출현하는 동시 출현빈도수를 측정하여 320×320 규모의 열을 나타낸 다음, 두 용어간의 유사도를 알기 위하여 식(1)을 이용하여 유사계수 값을 측정하였다. 이때 나온 값은 소숫점

〈표 1〉 용어의 동시출현 빈도수 행렬(n=25)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. BGT	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2. DAP	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3. KWIS	1	1	5	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4. MMPI	1	1	3	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5. D/X	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6. D/X관계	0	0	0	0	0	2	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7. P/O	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8. P/O관계	0	0	0	0	0	2	0	6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9. P/X	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10. P/X관계	0	0	0	0	0	1	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11. RPI	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12. TEST OF IRRELEVANCE	0	0	0	0	0	0	0	0	0	0	0	6	2	0	0	0	0	0	0	0	0	0	0	0	0
13. THUPSTONE	0	0	0	0	0	0	0	0	0	0	0	2	4	0	0	0	0	0	0	0	0	0	0	0	0
14. U형곡선	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
15. Y형미로	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	2	0	0	0	0	0	0
16. 모순불일치의 태도	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
17. 망상형	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
18. 미래행동	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
19. 무력감	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	7	0	0	0	0	0	0
20. 물질적 보상	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
21. 문항신뢰도지수	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
22. 우선적	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	2
23. 총체적 형태	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
24. 총체지향적 약효확	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
25. 청각	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	3

40) C.C. Gotlieb and S. Kumar, "Semantic...," pp. 495~497.

〈표 2〉 용어의 유사계수 행렬(n=25)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. BGT	1.000	1.000	.200	.125	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
2. DAP	1.000	1.000	.200	.125	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
3. KWIS	.200	.200	1.000	.300	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
4. MMPI	.125	.125	.300	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
5. O/X	.000	.000	.000	.000	1.000	.000	1.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
6. O/X관계	.000	.000	.000	.000	.000	1.000	.000	.286	.000	.250	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
7. P/O	.000	.000	.000	.000	1.000	.000	1.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
8. P/O관계	.000	.000	.000	.000	.000	.286	.000	1.000	.000	.143	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
9. P/X	.000	.000	.000	.000	1.000	.000	1.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
10. P/X관계	.000	.000	.000	.000	.000	.250	.000	.143	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
11. RPI	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
12. TEST OF IRRELEVANCE	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.250	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
13. THURSTONE	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.250	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
14. U형곡선	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
15. Y형미로	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
16. 모순불일치의 태도	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
17. 망상형	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000
18. 미래행동	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000
19. 무력감	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000
20. 물질적 보상	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000
21. 문항신뢰도 지수	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000
22. 우선적	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000
23. 총체적 형태	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000
24. 총체지향적 약효과	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
25. 청각	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000

4째 자리에서 반올림하여 $n=320$ 구모의 행렬로 표시하였다.

다음에 보여주는 두개의 행렬은 이미 형성된 용어의 빈도수와 동시 출현 빈도수를 나타내주는 행렬과 그 값에 유사도가 계산된 유사계수 행렬로 전체 행렬 중 $n=25$ 규모의 행렬만 나타낸 것이다.

3.2 용어군 형성

앞장에서 언급한대로 두 용어간의 유사성은 유사도 측정공식에 의하여 알 수 있으나 실제로 두 용어간의 유사성여부의 판정은 유사성크기를 나타내는 유사계수와 선택된 일정 기준치와의 비교로 이루어진다. 즉 유사계수가 기준치보다 클 때 비로서 두 용어가 유사하다고 판정내릴 수가 있다. 그러므로 용어군을 형성하기 전에 유사성 판정을 위한 기준치 T값을 결정해야 한다. 또한 T값은 형성될 용어군의 수를 필연적으로 결정하므로, 기준치 T는 만들려는 용어군 성격에 맞게 선택되어야 한다.

본 실험에서는 기준치를 0.2와 0.3으로 선택하였고, 먼저 부분연결 그래프를 형성하기전에 그래프 형성을 용이하게 해주는 이원행렬을 각각 기준치에 따라 두개 만들었다. 즉 유사계수가 선택된 기준치 값(0.2와 0.3)보다 크거나 같을 때

〈표 3〉 기준치 0.2로 형성된 이원행렬($n=25$)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. BGT	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2. DAP	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3. KWIS	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4. MMPI	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5. O/X	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6. O/X과계	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7. P/O	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8. P/O관계	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9. P/X	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10. P/X관계	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
11. RPI	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12. TEST OF IRRELEVANCE	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
13. THURSTONE	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
14. U형곡선	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
15. Y형미로	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
16. 모순불일치의 태도	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
17. 방상형	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
18. 미래행동	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
19. 무력감	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
20. 물질적 보상	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
21. 문항신뢰도 지수	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
22. 우선적	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
23. 총체적 형태	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
24. 총체지향적 약호화	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
25. 청각	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1

〈표 4〉 기준치 0.3으로 형성된 이원행렬($n=25$)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. BGT	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2. DAP	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3. KWIS	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4. MMPI	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5. O/X	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6. O/X관계	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7. P/O	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8. P/O관계	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9. P/X	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10. P/X관계	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11. RPI	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12. TEST OF IRRELEVANCE	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
13. THURSTONE	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
14. U형곡선	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
15. Y형비로	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
16. 모순불일치의 태도	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
17. 망상형	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
18. 미래행동	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
19. 무력감	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
20. 물질적 보상	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
21. 문항신뢰도 지수	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
22. 우선적	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
23. 총체적 형태	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
24. 총체지양적 악호화	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
25. 청각	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1

1이라는 값을 주고 그와 반대로 작을 때 0라는 값을 주어 320×320 규모의 이원행렬 두개를 만들었다.

다음의 표3과 표4는 이원 행렬 중 $n=25$까지의 부분을 발췌한 것으로서 표3는 기준치 0.2로, 표4는 기준치 0.3을 기준으로 하여 만들어진 것이다.

〈표 5〉 부분연결 그래프방식으로 형성된 용어군1

기준치	용어군 1에 속한 용어
0.2	BGT DAP 로샤검사 한국판단축형 KWIS MMPI 전환증 검사배터리 신경증 개별검사
0.3	BGT DAP 한국판단축형 로샤검사

용어군의 형성은 만들어진 이원행렬을 이용하여 두번 시행되었으며 그 결과 기준치 값이 0.2인 경우에는 용어군에 속한 용어의 총수는 300개이며 형성된 용어군의 수는 37개인 것에 반해, 0.3인 경우는 용어군의 수는 50개이고 총 용어의 수는 292개이다.

표 5는 부분연결그래프 방식으로 형성된 용어군 중 첫번째 용어군을 보여준 것이다.

본 실험에서의 부분연결 그래프방식은 용어실험집단을 크게 나누는데 그 목적이 있으므로 확대개념용어군을 형성하는 용어군 재형성 실험에

서는 좀 더 많은 용어가 연관되어 있는 0.2를 기준치로 한 용어군을 대상으로 삼았다.

3.3 용어군 재형성

용어군의 재형성은 두 단계로 이루어진다. 먼저 용어군에 속하는 용어간의 유사계수가 모두 기준치 값보다 크거나 같은 용어들로만 모아지게 하는 완전연결 그래프 방식으로 핵심개념용어군을 이미 형성된 용어군에 따라 형성하였다.

표 6은 BGT를 기준으로 삼아 형성된 첫번째 용어군과 이를 다시 핵심개념용어군으로 나타낸 것이다.

〈표 6〉 용어군 1과 핵심개념용어군

용어군 1	BGT DAP 로샤검사 한국판단축형 KWIS MMPI 전환증 검사배터리 신경증 정신병 개별검사
핵심개념용어군	1. BGT DAP 로샤검사 한국판단축형 2. BGT DAP 로샤검사 KWIS MMPI 3. BGT DAP MMPI 한국판단축형 전환증 4. 로샤검사 KWIS MMPI 5. 로샤검사 한국판단축형 KWIS MMPI 검사배터리 신경증 정신병 6. 한국판단축형 전환증 7. MMPI 검사배터리 개별검사 8. MMPI 신경증 정신병 9. 검사배터리 개별검사

다음 단계는 실제 검색에 이용될 수 있는 검색용 디소오러스를 작성하는 단계로 부분연결 그래프 방식으로 형성한 용어군에 있는 용어보다 더욱 유사한 관계가 있는 용어만을 포함하는 작은 규모의 용어군으로 형성하나, 핵심용어군보다는 큰 규모의 용어군인 확대개념 용어군을 형성하는 단계이다. 확대개념용어군은 고틀리브와 쿠머가 제시한 식(2)을 이용하여 형성한다.

본 실험에서는 0.2를 기준치로 한 용어군을 대상으로 확대개념용어군을 형성하였다. 핵심개념용어군을 결합시키는 기준치는 0.5, 0.6, 0.7, 0.8 모두 선택하였으며 마지막으로 부분연결 그래프 방식으로 형성된 용어군에 속한 용어가 핵심개념용어군의 결합을 통하여 전부 모아지는 값도 제시해 주었다[41].

4. 실험결과

4.1 용어군의 특성에 관한 분석

본 항에서는 기준치 T값에 따른 용어군 형성의 변화에 대해서 살펴보았다. 즉 3.1절에서 만들어진 유사계수 행렬에 기준치 0.2와 0.3을 적용시켜 만들어진 두개의 용어분류를 용어의 크기, 용어의 수 등을 기준으로 하여 분석해 보았다.

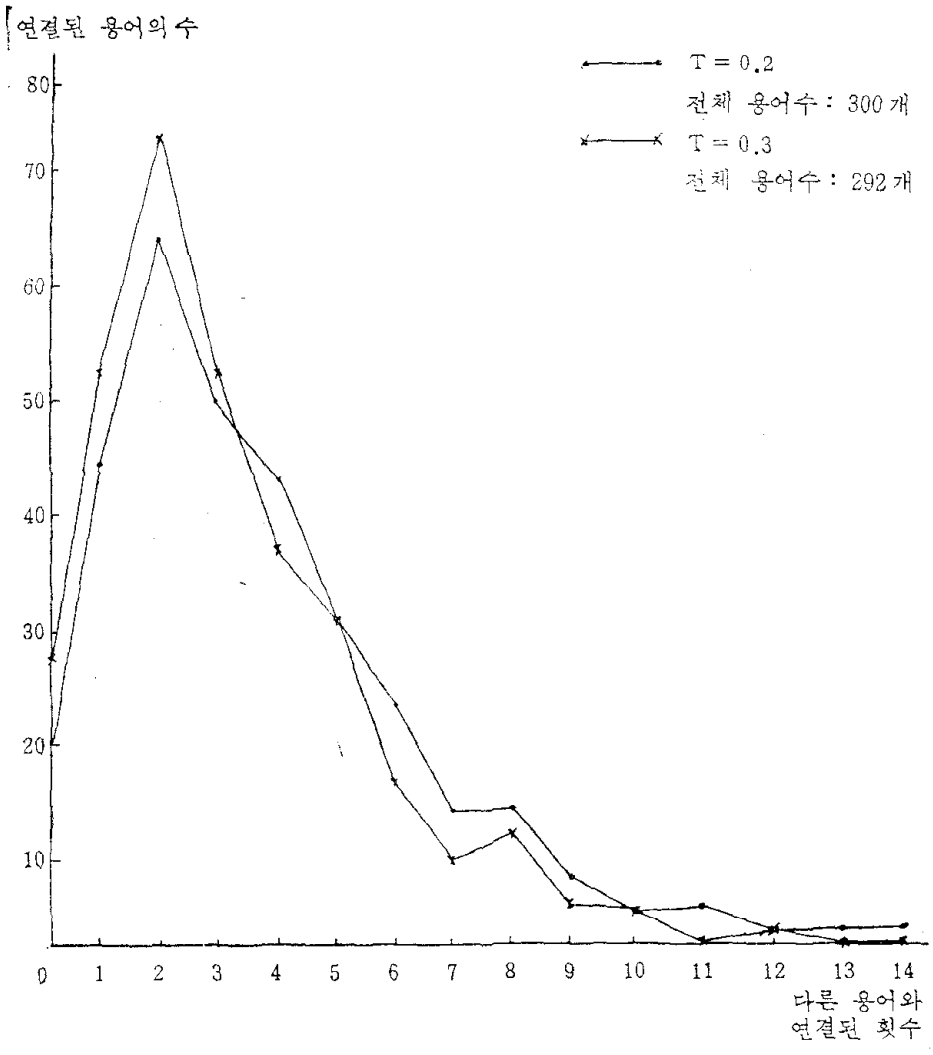
앞절에서 형성된 용어군을 조사해본 결과 기준치 T값의 증가에 따라 한 용어가 다른 용어와 연결되는 횟수가 감소됨과 동시에 추출되는 색인어의 수도 줄어들었음을 알 수 있었다. 즉 T값이 0.2인 경우에 추출된 색인어의 수는 300개인 것에 반에 0.3인 경우에는 292로 8개의 용어가 줄었다. 그와 반대로 형성된 용어군의 수는 T값의 증가에 따라 37개에서 50개로 늘어났음을 보여주었다. 그러나 이 결과는 비교 기준치 값 차이가 크지 않아서 생겨난 결과로 실제 0.2와 0.7를 비교한다면 역시 용어군의 수도 줄어들 것이다[42].

다음은 기준치 값에 따라 한 용어가 다른 용어와 연결된 빈도수를 살펴본 결과 그림 3과 같이 T가 0.2인 경우가 0.3인 경우보다 더 많이 다른 용어와 연결되고 있음을 보여주고 있다.

다음에 보이는 그림 4는 한 용어군에 포함된 용어의 수를 나타낸 것으로 평균 용어군의 크기

41) 용어군 재형성의 완전한 실험결과는 부록참조

42) 오거스톤과 인커의 실험에서 0.4와 0.7을 비교한 결과, 0.4인 경우에는 용어군이 402개가 생긴 반면에 0.7인 경우에는 148개로 현저하게 줄어들었다.



〔그림 3〕 한 용어와 연결된 용어의 수의 기준치 값에 따른 변화

도 기준치 값에 따라 변화된 것을 알 수 있다. 이 변화는 기준치 값의 변화에 따른 용어군의 수의 변동보다는 미미한 차이를 나타내고 있으며 큰 규모로 용어군을 형성하여 비교하는 실험에서도 기준치 값에 따른 용어군의 크기가 크게 변화되지 않음이 입증되었다[43].

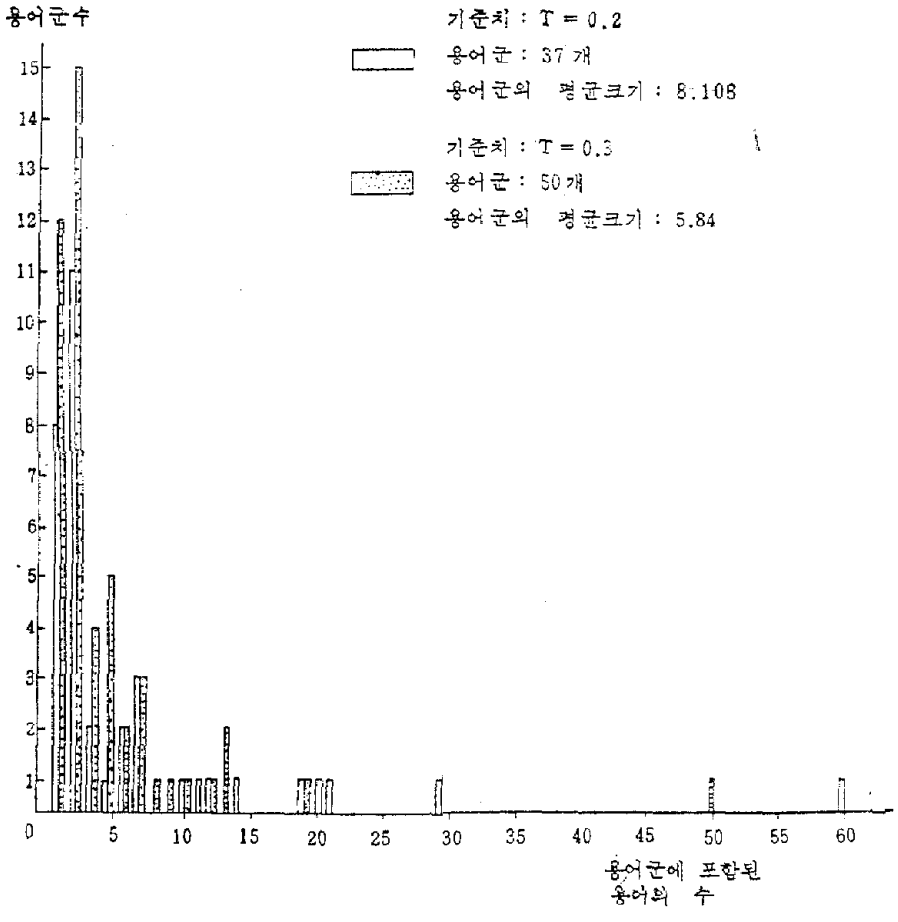
이상과 같은 분석을 종합해보면 용어군을 형성하는데 있어서 기준치는 한 용어군에 포함되는 용어의 수(용어군의 크기)에는 크게 영향을 미치지 않으나, 용어군을 형성하는 전체 용어의 수 및 형성되는 용어군의 총수에는 직접적인 관련이 있음을 알 수 있다.

4.2 용어군 내용에 관한 분석

본 항에서는 최종적으로 형성된 확대개념용어군에 속하는 용어의 어의적인 관계를 살펴보았다.

표5에서 보여주고 있는 용어군1(기준치 0.2로 형성된 용어군)은 부분연결 그래프 방식으로 형성된 용어군으로서, 용어군 재형성 단계에서 9개의 핵심개념용어군으로 나뉘어진 다음(표6 참조) 다시 5개의 확대개념용어군(기준치가 0.5인 경우)으로 재형성되었다. 이와 같이 재형성된 확대개념용어군에 속한 용어의 어의적 관계를 살펴보면 부분연결 그래프방식으로 형성된 용어군에서

43) 오거스톤과 민커의 실험에서 3,950개의 용어로 기준치 0.4와 0.7로 용어군의 형성한 결과 0.4인 경우 용어군이 평균 크기가 5.11이고 0.7인 경우 5.16으로 나타났다.



〔그림 4〕 기준치 0.2와 0.3에 따른 용어군의 구성

나타난 어의적 관계와 별 차이점이 없음을 알 수 있었다. 즉 용어군 1에 나타난 두개의 개념—임상진단검사와 정신장애병명—을 의미하는 용어가 확대개념용어군에서도 서로 분리되어 나타나지 않았으며, 단지 밀접한 유사관계가 있는 용어들이 어우러져서 형성되었을 뿐이다. 이로써 부분연결 그래프 방식으로 형성된 용어군에 속하는 용어가 이미 어의적 관계가 있는 용어들로 구성되었다면 확대개념용어군에서도 용어의 어의적 관계가 성립된다는 것을 알 수 있으며, 또한 용어군 재형성 실험은 더욱 유사한 용어들만을 구성하는 용어군의 형성방법이지, 실제 개념을 분류하는 방법이 아님이 밝혀졌다.

이 실험결과에서 나타난 또하나의 중요한 결과는 용어군이 거의 관련어로 구성되어 있다는 사실이다. 즉 동의어가 나타난 용어군은 세군데[44] 이고 그외에는 거의 관련어들로 구성되어 있으며 특히 한 용어에 대한 반대개념 및 비교개념이 비교적 잘 나타나 있었다.

그러나 본 실험결과인 확대개념용어군에는 어의적 관계가 없는 용어들로 형성된 용어군이 10% 정도 있었다. 이것은 분류방법에서 기인한 것이 아니라 실험집단을 구성할 때 적용한 복합어 통제와 불용어 선정이 미흡해서 나온 결과이다.

또한 본 실험의 표본수의 빈약성과 초록 구성상의 문제로[45] 어의적으로 밀접한 관계가 있는 용어들이 따로 분리되어 자각의 용어군으로 형성

44) 용어군 9의 {내적요인, 내적동기…}, 용어군 12의 {융화, 동화…}, 용어군 26의 {영속성, 항상성}

45) 초록에서 핵심개념을 나타내는 용어는 일반적으로 앞에서만 서술되고 다음 문장에서는 대명사로 서술되거나 생략됨으로서 용어가 동시출현을 못하여 나타난 현상이다.

된 경우도 발견할 수 있었다.

지금까지 확대개념용어군에 나타난 어의적-관계 및 그 문제점을 살펴보았다. 그 결과 용어의 통계적 속성에 근거하여 시행되는 용어자동분류 방법은 용어를 분류시켜 용어가 나타내는 각각의 주제를 세분화시키는 것이 아니라 어의적으로 관련이 있는 용어들을 모으는 방법이므로 큰 규모의 디소오러스 작성 방법에 용어자동분류가 효과적이라는 결론을 내릴 수 있다.

Ⅳ. 자동분류의 응용

용어의 자동분류는 자연어검색시스템에 있어서 검색시에 생기는 검색 실패를 최대한으로 줄여주는 검색용디소오러스를 구성하는 한 방법이며, 궁극적으로 검색효율을 높히는데 그 목적이 있다. 본 장에서는 자동분류가 검색시스템에서 실제로 어떻게 응용될 수 있는가를 살펴보고자 한다.

1. 디소오러스 작성

정보검색시스템에서 정보는 통제된 언어와 통제되지 않는 자연어를 사용하여 검색할 수 있다. 검색어에 대한 여러 연구결과를 보면 통제된 어휘를 사용한 경우가 자연어로 한 경우보다 검색효율이 현저히 높음을 알 수 있다. 디소오러스에 의해서 검색하는 것은 통제된 언어에 의한 검색으로서 검색용디소오러스는 색인자와 검색자 사이에 위치하여 동일한 주제의 용어를 선별해 주고 연결해 주어 검색실패를 미연에 방지하는 역할을 한다.

본 실험에서 형성된 용어자동분류는 디소오러스를 구성하는 한 방법으로서, 전문가의 지적인 작업을 완전히 대신해준다기보다는 다만 디소오러스 구성에 필요한 전문가의 노력을 도와주는데 그 의미가 있다고 본다. 통계적 정보는 어휘평가에 있어서 인간의 판단을 앞서지는 못하지만 평가하는데 유용한 근거가 될 수 있기 때문이다. 다음은 디소오러스를 구성하는데 용어자동분류가 응용되는 분야를 살펴본 것이다.

첫째, 용어자동분류는 동의어와 관련어의 선정 및 디스크립터 추출방법으로 응용될 수 있다. 용어자동분류를 우리말에 적용한 결과 형성된 용어군내에는 한 용어와 의미적 관계가 있는 모든 용어가 모아져서 나타나므로 동의어 및 관련어를 쉽게 찾아볼 수가 있다.

또한 같은 의미를 나타내는 용어 중에서 다른 용어보다 용어출현 빈도수가 높은 용어가 이용자에게 쉽게 접근될 수 있다는 가설아래 디스크립터는 쉽게 선정될 수가 있다.

둘째, 용어자동분류는 디소오러스 작성에 대한 기초연구로서 선행된다. 예를 들면, 여러 주제범위에 속하는 용어들을 자동분류하여 똑 같은 용어가 각 주제에 따라 어떻게 사용되며, 그 개념이 어떻게 변하는 가를 조사하는 연구와 디소오러스 크기, 형성된 용어군의 수, 용어군에 속하는 용어의 수를 조절하여 검색효율을 비교하는 연구 등을 들 수 있다. 이로써 이런 연구를 기초로 하여 실제 디소오러스를 작성할 때 가장 적합한 디소오러스를 구성할 수 있게 된다.

2. 검색어의 확장

만약 색인이 적합한 문헌을 찾을 수 있는 기회를 확장시키는 재현률향상수단과 이용자요구와 적합하지 않는 문헌을 배제시키는 정확률향상수단으로 묘사된다면 본 실험에서 응용된 용어자동분류는 재현률을 높이는 도구로 볼 수 있다[46].

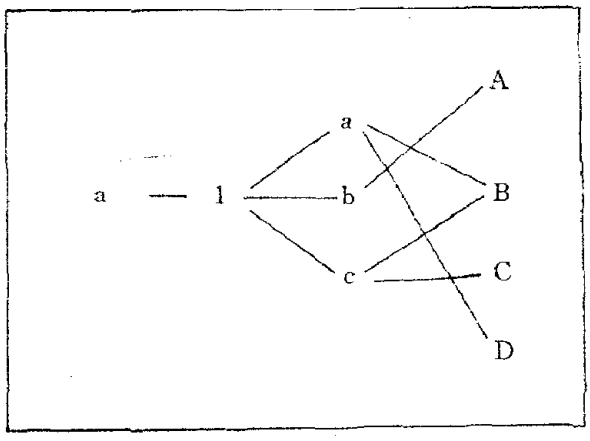
46) Karen Sparck Jones, *Antomatic...*, pp. 11~12.

즉 용어자동분류는 검색자가 질문에 나타난 용어로만 검색할 때 찾지 못한 적합한 문헌을 검색시 검색어를 확장하여 찾을수 있도록 하는 방법이기 때문이다.

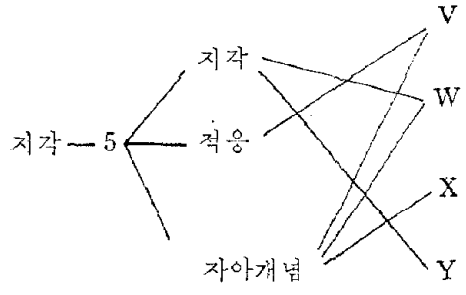
정보검색시스템에서 용어자동분류를 통하여 검색어가 확장되는 방법을 유형별로 나누어 보면 다음과 같다. 각 유형은 그림으로 표시하였으며, 그림에서 쓰여진 알파벳 소문자는 용어를 나타내며 특히 왼쪽에 쓰여진 알파벳 소문자는 검색자의 요구분에 나타난 용어이고, 숫자는 용어군, 알파벳 대문자는 문헌을 나타낸다. 각 유형별 예는 용어군 26을 대상으로 살펴보았다.

첫번째 유형은 가장 기본적인 방법으로 탐색하는 용어는 하나이고 이 용어가 속한 용어군을 이용하여 검색어를 확장시키는 방법이다.

유형 1

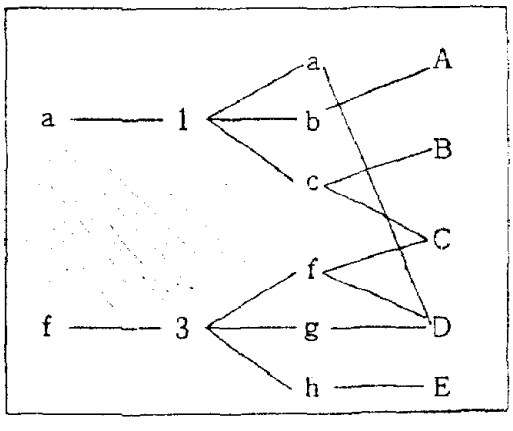


예 1

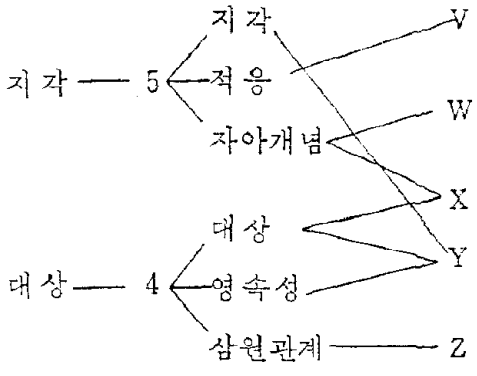


둘째로 검색하려는 용어가 두개 이상인 경우이다.

유형 2



예 2



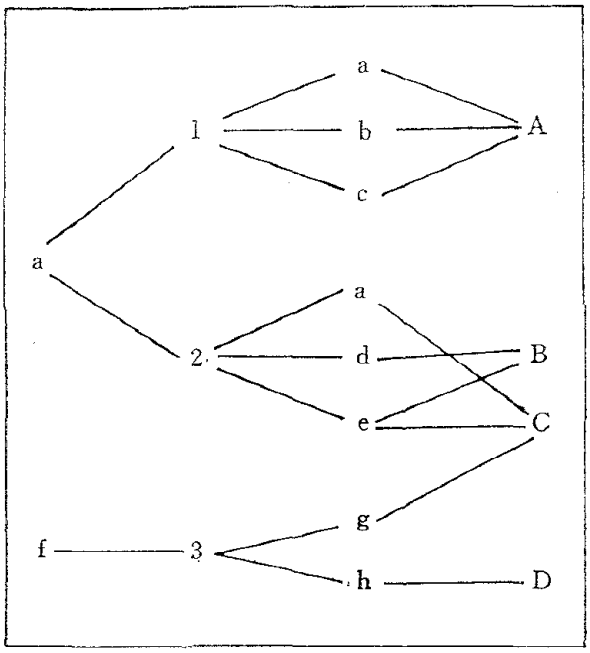
이 경우도 유형 1과 마찬가지로 두개의 용어를 모두 포함하는 문헌의 검색은 물론, 이 두 용어가 속한 용어군 내의 다른 용어들도 이용하여 폭넓은 검색을 할 수 있다. 즉 용어 a. f. g가 나타난 문헌 D와 용어 c. f가 같이 나타난 문헌 C도 검색할 수 있다.

47) V : 자아개념과 적응에 대한 연구
W : 부모의 태도와 자녀의 자아개념 형성간의 상관연구
X : 자아개념과 구성요인에 대한 연구
Y : 지각항상성과 대상영속성 개념의 발달에 관한 발생적 고찰

48) Z : 삼원적 사회관계 지각에서의 타인과 대상의 구체성, 균형 및 성의 영향

세번째 유형은 검색자가 a. f 두 검색어로 질문하여 검색한 결과 용어 a. f가 동시에 나타난 문헌은 없고 용어 a만 포함하는 문헌만 있는 경우이다.

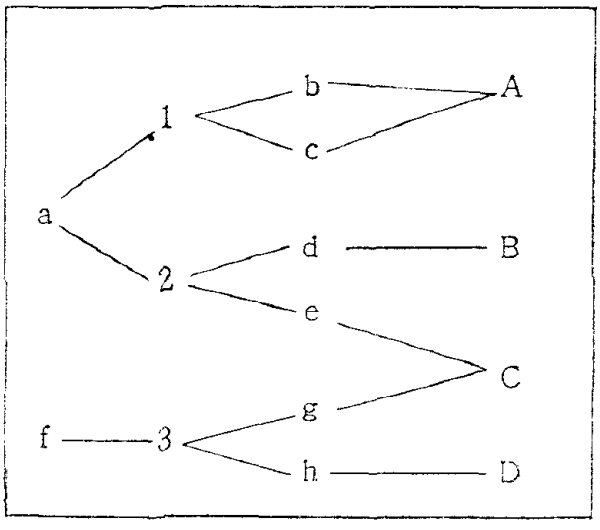
유형 3



이 경우에는 f의 유사어 g. h를 이용하여 검색을 확장시킬 수 있다. 즉 f의 유사어 g와 a가 함께 출현한 문헌 C를 검색할 수 있다.

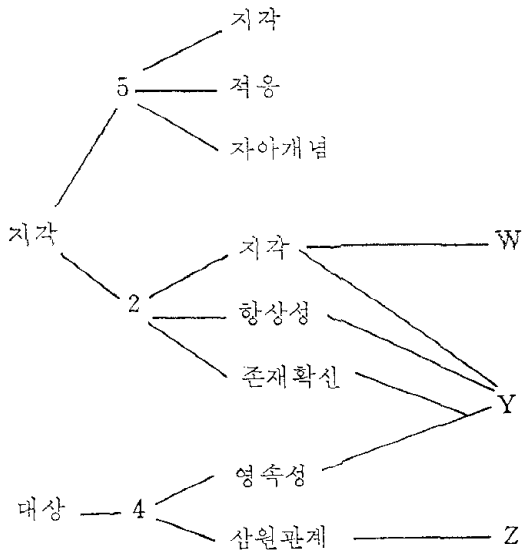
다음 유형 4는 검색자가 요구하는 용어 a와 f가 나타나는 문헌이 없는 경우이다. 이 경우에도 용어 a의 유사어와 f의 유사어를 이용하여 문헌 C를 검색할 수 있다.

유형 4

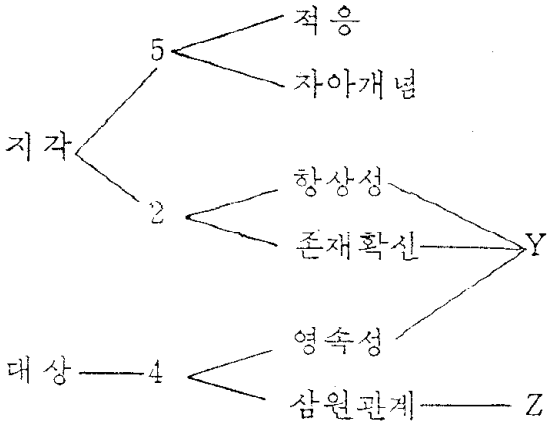


이와 같이 용어의 자동분류는 검색시 검색어로 확장시킬 수 있는 상호 대치어군을 형성하여 검색자가 더욱 편리하게, 그리고 오차없이 검색할 수 있게 하는 방법으로 사용되어 궁극적으로 정보검색효율을 크게 향상시킨다.

예 3



예 4



V. 결 론

최근에 자연어 정보검색시스템의 이용이 점차 늘어나게되자 이에 따라 시스템의 정보검색효율

을 높여주기 위한 검색용 디소오러스 자동작성에 대한 연구가 많이 진행되어 왔다.

본 논문에서는 검색용디소오러스를 작성하는 한 방법으로 자연어의 통계적 특성이 각 용어간의 어의적 관계를 나타낸다는 가설을 근거로 하는 용어자동분류를 선택하여 우리말에 적용시켜 보았다. 심리학 분야의 용어에 실험해 본 결과 37개의 용어군안에 총 99개의 확대개념용어군을 형성하였다.

그 결과에 따라 다음과 같은 결론을 내릴 수 있다.

1. 용어군을 형성하는데 있어서 기준치 값은 용어군 평균크기와는 무관하다 용어군을 형성하는 전체 용어의 수 및 형성된 용어군의 총 수에는 직접 관련하므로 만들려는 디소오러스의 규모 및 성격에 따라 기준치 값을 선정하는 것이 유용하다.
2. 부분연결 그레프 방식으로 형성된 용어군에 나타난 용어의 어의적 관계는 용어군이 재형성되어도 계속 유지되므로 일차적 용어군 형성에 유의하는 것이 바람직하다.
3. 용어자동분류는 용어가 나타내는 주제에 따라 세분화시키는 것이 아니라 단지 어의적으로 관련있는 용어를 한 용어군에 모으는 방법임으로 대규모 디소오러스 작성에 효과적으로 응용된다.
4. 용어자동분류로 형성된 용어군에 일반적인 색인이나 사전에 밝혀있지 않는 관련어들이 서로 밀접하게 연결되어 나타나므로, 특정 주제의 전문적인 디소오러스를 작성할 때 관련어 선정에 도움을 줄 수 있다.
5. 확대개념용어군은 검색시 용어의 상호대치가 가능한 용어들의 집합으로서 정보요구에 대응해줄 수 있는 검색어는 최대한 용어군에 나타난 용어의 수 만큼 확장될 수 있다. 이와 같은 검색어 확장으로 재현률이 향상되며, 또한 정보검색효율이 높아지게 된다.

REFERENCES

1. 사공철, 「정보검색에 있어서의 Thesaurus 도입에 관한 연구」 연세대 산업대학원 석사논문, 1974.
2. _____, 「정보검색론」 서울 : 아세아문화사, 1977.
3. 이정일, "Thesaurus의 이용과 최근의 방향," 정보관리연구, 14권, 2호(1981. 6), pp. 69-75.
4. 정문성, 「정보처리체계에서 Clustering」 과학원 석사논문, 1975.
5. 정영미, "Document Clustering에 의한 정보검색," 경영과 컴퓨터 5권, 6호(1980. 6), pp. 52-58.
6. 최운도 등저, 「교육 Thesaurus 개발연구」 서울: 한국교육개발원, 1981.
7. 현은정, "정보검색 시스템과 어휘조정," 정보관리연구, 13권, 4호(1980. 8), pp. 119-123.
8. Auguston, J.G. and J. Minker, "Deriving Term Relationship for a Corpus by Graph Theoretical Clusters," *JASIS*, Vol. 21, No. 2 (March, 1970), pp. 101-111.
9. _____, "An Analysis of Some Graph Theoretical Cluster Technique," *J. of ACM*, Vol. 17, No. 4 (Oct., 1970), pp. 571-588.
10. Bonner, R.E., "On Some Clustering Technique," *IBM Journal of Research and Development*, Vol. 8, No. 1 (Jan., 1964), pp. 22-32.
11. Borko, H. and M. Bernick, "Automatic Document Classification," *J. of ACM*, Vol. 10, No. 2 (April, 1963). pp. 151-162.
12. _____, "Automatic Document Classification," *J. of ACM*, Vol. 11, No. 2 (April, 1964), pp. 138-151.
13. Buchanan, Brian, *Theory of Library Classification*. London: Clive Bingley, 1979.
14. Dale, A.G. and N. Dale, "Some Clumping Experiments for Associative Document Retrieval," *American Documentation*, Vol. 16, No.

1 (Jan., 1965), pp. 5-9.

15. Dattola, R.T., "A Fast Algorithm for Automatic Classification," *J. of Library Automation*, Vol. 2, No. 1 (March, 1969), pp. 31-48.
16. Dillon, Martin, "Automatic Classification of Harris Survey Questions: An Experiment in the Organization of Information," *JASIS*, Vol. 33, No. 5 (Sep., 1982), pp. 294-301.
17. Dillon, M. and D. Caplan, "A Technique for Evaluating Automatic Term Clustering," *JASIS*, Vol. 31, No. 2 (March, 1980), pp. 89-96.
18. Doyle, L.B., "Indexing and Abstracting by Associative," *American Documentation*, Vol. 13, No. 4 (Oct., 1962), pp. 378-390.
19. Doyle, L.B., "Is Automatic Classification a Reasonable Application of Statistical Analysis of Text?," *J. of ACM*, Vol. 12, No. 1 (Jan., 1965), pp. 17-37.
20. Everitt, Brian, *Cluster Analysis*. New York: John Wiley and Sons, 1974.
21. Ghose, Amitabha and Anand S. Dhawle, "Problems of Thesaurus Construction," *JASIS*, Vol. 28, No. 4 (July, 1970), pp. 211-217.
22. Gotlieb, C.C. and S. Kumar, "Semantic Clustering of Index Term," *J. of ACM*, Vol. 15, No. 4(Oct., 1968), pp. 493-513.
23. Hoyler, W.G., "Automatic Indexing and Generation of Classification Systems by Algorithm," *Information Storage and Retrieval*, Vol. 9, No. 4 (April, 1973), pp. 233-242.
24. Kent, Allen and Harold Lancour ed., *Encyclopedia of Library and Information Science*. New York: Marcel Decker, 1969: In Vol. 2 "Automatic Analysis," by Mary Elizabeth Sterens.
25. _____, In Vol. 5 "Clumps, Theory of," by Karen Spark Jones.
26. Lancaster, F.W., et al., "Evaluating the Efficientiveness of On-Line Natural Language Retrieval System," *Information Storage and Retrieval*, Vol. 8, No. 5 (Oct., 1972), pp. 223-245.
27. Lancaster, F.W., "MEDLAS: Report on the Evaluation of its Operating Efficient," *American Documentation*, Vol. 20, No. 2 (April, 1969), pp. 119-142.
28. Leferver, M., Barbara Freedmen and Louis Schultz, "Managing an Uncontrolled Vocabulary Ex Post Facto," *JASIS*, Vol. 23, No. 6 (Nov., 1972), pp. 339-342.
29. Lesk, M.E., "Word-word Association in Document Retieval System," *JASIS*, Vol. 20, No. 1 (Jan., 1969), pp. 27-29.
30. Long, J.M. et al., "Dictionary Buildup and Stability of Word Frequency in a Specialized Medical Area," *JASIS*, Vol. 18, No. 1 (Jan., 1967), pp. 21-25.
31. Needham, R.M., "Application of the Theory of Clumps," *Mech. Transl. Comp. Linguist*, Vol. 8 (1965), pp. 113-127.
32. Peter Willet, "A Fast Procedure for the Calculation of Similarity Coefficients in Automatic Classification," *Information Processing and Management*, Vol. 17, No. 2 (Feb., 1981), pp. 53-60.
33. Rogers, D. and T. Tanimoto, "A Computer Program for Classifying Plants," *Science*, Vol. 132, No. 3434 (Oct., 1960), pp. 1115-1118.
34. Salton, Gerald, *Automatic Information Organization Retrieval*. New York: MCH., 1968.
35. Salton, Geral, *The SMART Retrieval System; Experiments in Automatic Document Processing*. New Jersey: Prentice-Hall, 1971.
36. _____, *Dynamic Information and Library Processing*. New Jersey: Prentice-Hall, 1975.
37. Salton, Gerald and C.S. Yang, "A Theory of Term Importance in Automatic Analysis," *JASIS*, Vol. 26, No. 1 (Jan., 1975), pp. 33-44.
38. Salton, Gerald and Michael J. McGrill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book Company, 1983.
39. Soergel, Dagobert, *Indexing Language and Thesaurus: Construction and Maintenance*. LA: Melville Publishing Co., 1974.
40. Sparck Jones, Karen, "Some Thoughts on Classification for Retrieval," *J. of Doume-*

ntation, Vol. 26, No. 2 (June, 1970), pp. 89-101.

41. _____, *Automatic Keyword Classification for Information Retieval.* London: Butterworth, 1971.
42. Sparck Jones, K. and D.M. Jackson, "Current Approaches to Classification and Clumpfinding at the Cambridge Language Research Unit" *The Computer Journal, Vol.* 10, No. 1 (Jan., 1967), pp. 29-37.
43. Sparck Jones, K. and D.M. Jackson, "The Use of Automatically Obtained Keyword Classifications for Information Retrieval," *Information Storage and Retrieval*, Vol. 6, No. 4 (August, 1970), pp. 175-185.
44 Sparck Jones, K. and E.O. Barber, "What Makes an Automatic Keyword Classification Effective," *JASIS*, Vol. 22, No. 3 (May, 19 71), pp. 166-175.
45. Stile, H.E., "The Association Factor in Information Retrieval," *J. of ACM*, Vol. 8, No. 2 (April, 1961), pp. 271-279.
46. Van Rijsbergen, C.J., *Information Retrieval* 2nd ed. London: Butterworths, 1979.
47. _____, "A Theoretical Basis for the Use of Co-occurence Data in Information Retrieval," *J. of Documentation*, Vol. 33, No. 2 (June, 1971), pp. 106-119.
48. Van Ryzin, J. ed., *Classification and Cluste ring*. New York: Academic Press, 1977.
49. Wallace, C.S. and D.M. Bouton, "An Information Measures for Classification," *The Computer Journal*, Vol. 11, No. 2 (May, 1968), pp. 185-194.
50. Yu, Clement T., "The Stability of Two Common Matching Functions in Classification with Respect to a Proposed Measure," *JASIS*, Vol. 27, No. 4 (July, 1976), pp. 248-255.
51. Yu, Clement T. and Vijay V. Raghavan, "Single-Pass Method for Determinating the Semantic Relationships between Terms," *JASIS*, Vol. 28, No. 6 (Nov., 1977), pp. 345-354.