정보 검색에 있어서 커어널 (KERNEL) 기법에 관한 연구

정 준 민*

<목 차>

Ⅰ. 서 언

오늘날 정보검색 시스팀을 이용하고 있는 이용자들에 있어서 가장 당면한 문제는 대규모의 데이터베이스 내에서 자신의 정보 요구를 충족시켜줄 수 있는 소수의 정보만을 어떻게하면 효율적으로 검색할 수 있을까 하는 일이다. 실제로 정보검색 시스팀을 사용하는 이용자들의 대부분은 그들이 이용하고 있는 데이터베이스의 구조와 시스팀에 대해선 거의 지식을 갖지 못하고 있는 실정이다.

정보검색 시스팀의 대부분은 그것이 전산화가 되었든 아니든간에 데이터베이스를 구성하고 있는 정보와 그것을 특징지울 수 있는 장치(분류번호, 색인어)를 그 정보에 부여함으로써 이용자들로 하여금 데이터베이스에 접근할 수 있도록 설계되어져 있다. 이용자들이 데이터베이스에 접근하기 위해서 가장 많이 사용되어지고 아직도 대부분의 시스팀이 채용하고

* 全南大學校 文獻情報學科

있는 정보검색 또는 탐색 기법 중의 하나가 부울(Boole) 대수의 논리 형식이다. 즉, 정보 이용자가 요구하는 질문을 복수의 탐색어로 표시하고 각 탐색어의 관계를 파악하여 그 관계를 논리적(論理積 ; AND), 논리화(論理和 ; OR) 또는 논리차(論理差 ; NOT)의 논리연산자로 조합하여 데이터베이스 내의 각 문헌을 탐색, 질문의 논리조합과 일치하는 문헌을 검색하도록 설계되어져 있다.

그러나 정보검색의 효율은 비록 그것이 같은 질문과 같은 논리연산자에 의해 탐색이 되었다 할지라도 실지 이용하는 이용자들의 욕구를 완전히 만족시킨다고 볼 수는 없다. 결국 전통적인 부울 대수의 문제점 내지는 효율성에 대한 논란이 제기되었으며[11,21] 그것을 요약하여 보면,

첫째, 논리조합에 나타난 각 탐색어들의 상대적 중요도가 전혀 고려되지않고 있으며, 둘째, 부울 대수의 특성상 논리조합과 완전히 일치하는 문헌만 검색이 되고 부분적으로 일치하는 문헌은 검색에서 제외되어지고 있다. 마지막으로 검색된 문헌들의 탐색어에 대한 비중의 크기 순으로 정리할 수 없다는 것이다.

결국 전통적인 부울 대수의 단점을 보완하고자 부울 대수의 논리연산자에 가중치를 부여하는 기법이 개발되었다. 그러나 이 방법은 데이터베이스 내의 문헌에 부여된 색인어와 탐색 질문에 사용된 색인어 양쪽 또는 어느 한쪽에 필요에 따라 가중치를 부여하여야 함으로써 그 가중치를 줄 때 생기는 주관적 판단과 지속적으로 데이터베이스를 유지할 때 야기되는 일관성이 큰 문제로 제기 되고있다.

한편, 위와같은 부울 대수의 비합리성과 비효율성의 지적과 함께 컴퓨터 하드웨어의 발전은 새로운 검색기법의 필요성을 인식하게 하였다. 일반적으로 정보검색이라함은 단지 데이터베이스 내의 문헌과 이용자의 질문과를 연결시켜주는 탐색업무를 의미하는데, 엄격히 말한다면 정보를 가공, 축적하는 방법과 축적된 정보간의 상관관계를 규명짓는 작업도 정보검색의 한 과정이라 말할 수 있다. 즉, 정보검색이란 정보의 축적과 탐색을 동시에 표현하는 말인 것이다. 이런 각도에서 보면 정보검색에 대한

연구는 단지 탐색업무를 개선하는데서 부터 자동색인, 분류이론[29,30,33,38]; 언어학[20,31]; 질문어 분석[1,2]; 인용문헌분석[8,15,26,2]; 계층적, 확률검색론[10,12,23,24,34,35]; 나아가 인공지능[28] 분야까지도 총망라되는 것이다.

컴퓨터 하드웨어의 발전과 함께 새로이 발전한 정보검색 기법으로 클러스터링 기법과 확률검색 기법이 있다. 확률검색 기법이라함은 질문과 문헌간의 관계를 확률적 개념으로 정의하고 상대적으로 높은 상관 관계를 갖는 문헌을 검색하는 기법을 말한다[18,35]. 한편, 클러스터링 기법은 데이터베이스 내에서의 문헌들을 여러개의 소규모 집합(CLUSTERS)으로 나누어 특정 질문에 대한 적합 문헌군을 탐색해 내는 방법으로 이것은 유사문헌을 같은 클러스터로 묶어둠으로써 가능케한다.[34]

클러스터링 기법은 크게 수평적 구조와 계층적구조로 나누어 볼 수 있는바 수평적 구조는 탐색시 전 문헌을 또는 각 클러스터의 대표 문헌을 비교함으로써 이루어지나 계층적 구조는 문헌들의 상관관계에 따라 계층적으로 분류하여 놓음으로써 전체 문헌 집단이 아닌 특정한 계열에 속한 문헌을 비교해 봄으로써 검색이 가능하다.

그러나 클러스터링 기법에서 문제시되는 것 중의 하나가 데이터베이스 내의 문헌을 여러개의 클러스터로 나누어주는 클러스터 알고리즘의 개발이라 할 것이다.[22,24,35,37]

이에 본고는 클러스터링 기법의 특성을 분석하여 보고 클러스터 알고리즘으로 커어널 기법[3]을 새로이 도입하여 봄으로써 좀 더 개선된 검색시스템을 설계하여 보기로 한다.

Ⅱ. 클러스터링 기법

클러스터링 기법을 논하기에 앞서 우리는 문헌과 질문, 문헌과 문헌의 관계를 규명할 적합도 또는 상관계수에 대해 알아야겠다.

적합도(Relevance ; 상관계수)란 매우 주관적인 평가 단위이다. 서로 다른 이용자들간엔 같은 질문에 의해 검색된 특정 문헌에 대해 서로 상이

한 적합성 판정을 내릴 수가 있다. 그러나 정보검색시스팀을 설계하고 자동화된 검색 기법을 개발함에 있어서 적합성의 개념은 객관적으로 평가될 필요가 있다. 이와같은 적합성의 객관적 정의는 쿠퍼(W. S. Cooper)[6]에 의해 발표되어졌다. 즉, 그에 의하면 적합성이란 개념은 비록 그것이 주관적 판단과 다소 일치하지 않더라도 '논리적 결정'이란 용어로 표현 가능하다는 것이다. 이것이 이른바 부울 대수의 논리연산자에 의한 표출인 것이다. 그러나 앞서 언급한바와 같이 부울 대수에는 나름대로의 단점을 갖고 있으며 특히 논리연산자 중 논리화(OR)를 사용할 경우 너무 많은 비적합 문헌이 함께 검색될 수 있으며 논리적(AND)을 사용할 경우는 적합한 문헌 일부가 검색되지 않을 우려가 있는 것이다. 이것은 부울 대수에서 문헌의 취급 및 탐색시 각 문헌이 독립적으로 질문에 대응되기 때문에 나타나는 현상이라고 볼 수 있다.

여기에서 고프만에 의해 부울 대수의 단점을 보완할 수 있는 정보검색 기법으로 클러스터링 기법 도입의 근거가 마련되어진다.[11] 클러스터링이라함은 일반적으로 전체 데이터베이스 내의 총문헌을 각기 두개의 문헌으로 된 쌍을 만들어 각각의 상관계수를 측정, 전체 집단을 여러개의 하부구조로 구분하는 작업을 말하며, 이것은 개발되어진 알고리즘에 따라 서로 다른 형태를 취할 수 있다.

클러스터링 기법이 정보검색의 기법으로 활용되는 데에는 다음 두가지 가설에 의해 가능하다. 즉,

A. 특정 질문에 주어진 검색 문헌군은 그들 문헌간의 상관계수에 의해 조직되어지며 비록 각기 문헌을 특정 질문과 1대1 대응하지 않더라도, 서로의 상관계수에 따라 연결된 문헌들의 각자의 상관계수 값이 특정한 값 이상을 유지하고 있으면 이들 문헌 집단은 특정 질문에 적합하다.[11]

B. 같은 집단에 속하는 유사한 내용의 문헌들은 같은 질문에 대해 같은 집단 내에 존재하지 않는 문헌들 보다는 더 적합하다. (클러스

터 가설)[37]

클러스터링 기법은 문헌간의 상관계수를 측정하여 그것이 특정 기준치를 상회할 경우 하나의 클러스터로 묶여질 수 있으며 그 상관계수를 측정하는 데에는 여러가지 기법이 있을 수 있다.[35]

다음은 정보검색에 널리 사용되어진 상관계수 공식으로 $|X|$와 $|Y|$는 각 문헌에 나타난 색인어의 갯수이며 $|X \cup Y|$는 두 문헌에 나타난 총 색인어의 갯수이며 $|X \cap Y|$는 두 문헌에 동시에 나타난 색인어의 갯수를 의미한다.

다이스계수 $\dfrac{2|X \cap Y|}{|X|+|Y|}$

자카드계수 $\dfrac{|X \cap Y|}{|X \cup Y|}$

코싸인계수 $\dfrac{|X \cap Y|}{\sqrt{|X| \cdot |Y|}}$

중복도계수 $\dfrac{|X \cap Y|}{\min(|X|, |Y|)}$

그러나 이와같은 상관계수 값은 두 문헌 X와 Y의 상관관계를 같게 취급하는 문제점을 갖고 있다. 예를 들어 문헌 X가 $\{A, B, D, F, G\}$의 색인어를 갖고 문헌 Y가 $\{A, B, C\}$의 색인어를 갖을 경우 두 문헌의 상관계수 S는 다음과 같다.

다이스계수 $S(X, Y) = \dfrac{2 \times 2}{8} = \dfrac{1}{2}$

자카드계수 $S(X, Y) = \dfrac{2}{6} = \dfrac{1}{3}$

코싸인계수 $S(X, Y) = \dfrac{2}{\sqrt{5 \cdot 3}} = \dfrac{2}{\sqrt{15}}$

중복도계수 $S(X, Y) = \dfrac{2}{\min(5, 3)} = \dfrac{2}{3}$

위에서 보듯이 X와 Y의 문헌간 상관관계가 서로 같은 것으로 표현되

어 실제로 X가 Y에 대해 또는 Y가 X에 대해 갖는 상관계수 즉, 적합성의 관계는 명확히 규정되지 못하고 있다. 이것을 위에 언급한 상관계수와 같은 형식으로 표현하면서도 각 문헌의 상대문헌에 대한 적합성으로 표현하여 보면

$$P(X,Y)=\frac{|X\cap Y|}{|X|}$$

즉, $P(X,Y)=\frac{2}{5}$ $\quad$ $P(Y,X)=\frac{2}{3}$

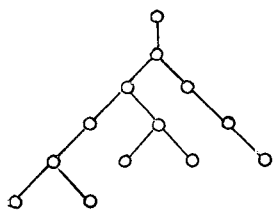
여기에서도 알 수 있듯이 문헌 Y가 문헌 X에 대해 더 많은 적합성을 부여하고 있다. 다시 말하면 문헌 Y를 검색하였을 경우는 문헌 X도 검색이 되나 문헌 X가 검색되었을 경우는 문헌 Y의 검색은 꼭 일어나지 않는다는 것이다. 물론 질문에 따라 문헌 X또는 Y의 선택이 달라질 것을 전제로 했을 경우이다.

이처럼 비대칭 상관계수는 고프만(W. Goffman)에 의해 제시되었으며[1] 데이터베이스 내에서의 문헌군을 비대칭 상관계수로 정의함으로써 포괄적이거나 많은 색인어를 탐색어로 했을 경우, 대칭형 상관계수를 사용했을 때 보다 좀 더 높은 정확률을 기할 수가 있다. 물론 어느 경우에든 가중치를 부여했을 경우는 또 다른 결과를 보여 줄 수도 있다.[24]

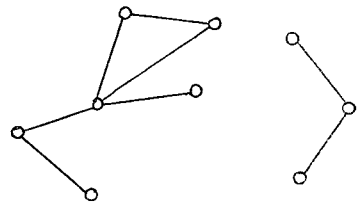
다음은 이상과 같은 상관계수를 통하여 구축된 데이터베이스의 상관계수 벡터를 이용하여 실지로 클러스터를 구축하여 검색 시스팀을 설계해 보기로 한다.

상관계수 벡터에 기초한 클러스터링 기법은 주로 그래프 이론을 이용하고 있으며 그래프의 형태적 특정에 따라 목(tree)구조, 부분 연결 그래프(connected graph), 완전 연결 그래프(maximal complete graph)로 나누어 클러스터의 형태를 결정한다.

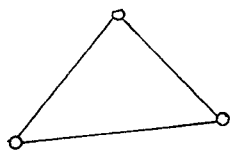
완전연결 그래프의 형태를 취하는 클러스터(clique)[9,40]에서는 같은 클러스터 내에 존재하는 문헌간에 상당히 높은 상관계수(적합성)를 갖고 있어 특정 질문에 대한 정확률이 높아지는 반면 완전연결형의 클러스터를

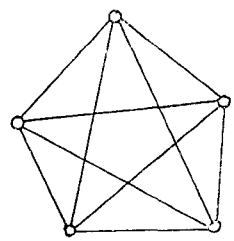


a. 목(tree) 구조



b. 부분연결 그래프





c. 완전 연결 그래프

생산해 내는데 많은 경비가 드는 단점이 있다. 한편, 다른 형태의 클러스터 구성에서는 물론 완전연결 클러스터에서도 문제가 되지만, 문헌과 문헌 사이의 상관계수가 특정 기준치(threshold)를 넘어야 하는데 그 기준치를 결정하는데 매우 큰 어려움이 있다. 이 문제를 객관적으로 해결하는 방법으로 'hyper-graph theory'[25]가 제시되고 있으나 아직은 많은 연구가 있어야 되리라 본다.

목(tree) 구조 형태를 취한 클러스터 화일에서 또한 문제가 되는 것은 클러스터의 상위 클러스터를 형성하거나 탐색시 각 클러스터를 대변하여 줄 센트로이드 문헌 또는 센트로이드 유사문헌(색인어 벡터로 구성된 가공의 문헌)이 만들어져야 하는데 이런 분석의 대표적 예가 패턴인식(pattern recognition) 기법의 검색 시스템 응용이다.[5]

이 과정을 설명하여 보면 다음과 같다.

1) 데이터베이스 내의 문헌을 임의로 k개의 그룹으로 나눈다. (여기서 k라 함은 데이터베이스 내에서 생성해낼 클러스터의 갯수를 의미

한다.)

2) 각각의 그룹내에서 각각 문헌에 부여된 색인어의 벡터 값의 거리개념에 따라 유형 또는 무형의 색인어 벡터(문헌 또는 유사문헌)을 찾아낸다.
3) 이렇게 산출된 k개 또는 그 이상의 문헌(cluster seed)을 중심으로 전체 데이터베이스 내의 문헌을 새로운 그룹으로 나눈다.
4) 새롭게 형성된 그룹내에서 2)의 과정을 되풀이하여 새로운 센트로이드 값을 찾아낸다.

이렇게 형성된 클러스터와 센트로이드는 그 데이터베이스를 나눌 수 있는 방법이 되며 앞에서 언급된 상관계수(similarity coefficient) 개념과는 다른 수용계수(cover-coefficient)라는 새로운 개념을 소개하고 있다. 즉 문헌 각자간의 상관계수 관계가 아닌 센트로이드를 중심으로한 문헌과 센트로이드간의 최적의(maximally) 수용관계를 갖게 되는 것이다. 특히, 새로 유입된 문헌이 어느 센트로이드에도 최적의 수용관계를 갖지 못할경우 단독으로 하나의 클러스터를 형성하게 되는 것이다.

계층적 클러스터 기법과 센트로이드 개념을 하나로 묶어 형성해낸 클러스터 기법으로 두 일본인 학자에 의해 개발되어진 2단계 클러스터링 기법이 있다.[19] 이것은 첫번째 단계에서 여려개의 클러스터가 구축되어지고 각 클러스터의 대표치로 센트로이드를 찾아내며, 둘째 단계에서 이들 센트로이드를 갖고 새로운 클러스터를 형성하여 상위 계층의 클러스터를 생성해 내는 기법이다.

도일(L. B. Doyle)[7]은 이미 이런 개념을 그의 '의미지도'를 만드는데 활용하고 있다. 도일에 의하면 데이터베이스 내에서 문헌들은 디소오러스의 구조와 비슷한 의미구조를 구성하고 있다고 보고 이용자는 그의 필요와 요구에 따른 방향으로 의미(색인어)를 선정 추적하여 가면 관련 문헌이 검색될 수 있다는 것이다. 이것은 색인어들을 통한 센트로이드 개념과 계층적 구조를 동시에 수용하고 있으며 이미 구축된 '의미지도(semantic

map)'위에 산재해 있으며 그 경로가 곧 이용자의 질문어(탐색어)의 집합으로 볼 수 있다. 리토프스키(B. Litofsky)[16]도 그 유사한 개념을 응용, 새로운 검색기법을 설명하고 있다.

이상의 클러스터링 기법을 종합하여 볼 때 다음과 같은 의문과 결과에 도달하게 된다.

1) 문헌과 문헌의 상관계수를 규정짓기 위한 객관적 방법은 무엇인가?
2) 각 클러스터를 형성할 때 그 형성의 근거가 되는 상관관계 기준치(threshold value)는 어떻게 결정할 수 있는가?
3) 형성된 클러스터로 부터 그 대표치(controid, key seed)는 어떻게 산출할 것인가?
4) 클러스터의 계층적 구조를 위한 구성은 어떻게 이룰것인가?

본고에서는 위의 4가지 문제점을 해결하기 위해 커어널(kernel) 기법을 도입하고 그것의 특성을 응용, 새로운 방법을 구축하여 보았다. 즉, 커어널을 형성하는 방향설정 그래프(Directed Graph) 모델을 통해 한 쌍이 된 두 문헌의 쌍방 동일한 상관계수 보다는 방향성을 갖는 상관계수(P)를 상관계수로 삼았으며 기준치 및 센트로이드 개념은 커어널의 특성에 의해 해결하였고 계층적구조는 커어널 집합에 속하는 문헌들을 새로운 클러스터로 구축 상위 개념의 커널(super kernel set)을 형성하여 하나의 문헌지도를 만듦으로써 해결하였다.

Ⅲ. 커어널 기법

순서쌍을 갖는 공집합이 아닌 집합 $V=\{v_1, v_2, \ldots v_n\}$ 와 그들의 순서쌍 $A \subset VXV$ 즉, (v_i, v_j)로 표시된 선으로 이루어진 집합 또는 방향성을 갖는 그래프 $G=(V, A)$가 있다고 할 때, 선(edge)으로 연결된 두점(u, v)에서 u는 v의 선임자로, v는 u의 후임자로 표시한다. 또한 요소 x와 부분집합 S가 그래프 V의 일부이며 ($x \subset V$, $S \subset V$), 그의 상관관계(Γ)가

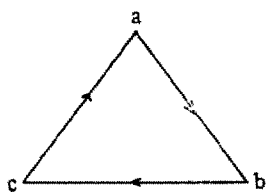
다음과 같이 정의되어있다면

$$\Gamma(x)=\{V:(x,\ v)\subset S\},\quad \Gamma(S)=\bigcup_{k\in s}\Gamma(k)^{*}$$

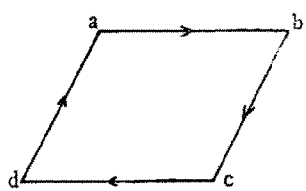
다음과 같은 조건을 만족하는 V의 부분집합 K를 커어널 또는 그래프 $G=(V,A)$의 핵심요소군이라 부른다.

$$\Gamma(u)\cap K=\phi,\quad \forall_u\in K \quad (1)$$
$$\Gamma(u)\cap K\neq\phi,\quad \forall_u\in \overline{K}\cap V \quad (2)$$

여기에서 이러한 점들의 집합 K는 (1)의 공식을 만족할 경우 안정성 또는 독립성이 있다고 하며 (2)를 만족할 경우 흡입성 또는 장악성이 있다고 표현할 수 있다. 이것을 그림으로 표현하면 다음과 같다.



커어널 집합이 없음



2개의 커어널 집합이 있음
(a, c), (b, d)

이상에서 보듯이 한 개의 그래프에서 커어널은 한개이상 또는 전혀 존재하지 않을 수도 있다.

클러스터링 기법이란 다차원 공간에서 각 노드(문헌)들을 선으로 연결, 같이 연결된 그룹내에 존재하는 노드끼리는 어떤 형태로든 상관관계를 유지하고 있으며 그룹 밖의 어느 노드들과는 다른 관계를 유지하게끔 데이터베이스 내의 노드들을 그룹으로 나누는 작업을 말한다. 전통적으로 표

* 요소 v가 (x,v)로 순서쌍으로 이루어져 있으며 이것이 부분집합 S에 포함될때 「$\Gamma(x)$란 이런관계에 있는 모든 v를 말하며 「(s)란 그런 관계에 있는 점들의 집합을 말한다.

준이 되는 상관관계란 각 그룹을 객관적으로 분류할 수 있는 것들이었으며 같은 그룹 내에서는 그 그룹의 갖는 공통의 의미나 속성(예 : 색인어)을 공유할 수 있어야 한다. 즉, 한개의 클러스터를 형성하기 위해 모인점들은 유사성에 있어서 객관성을 갖어야 하며 다른 클러스터의 점들과는 다른 속성을 나타내어야만 하는 것이다.

클러스터링 기법에서 가장 어려운 점은 모든 점들이 특정 클러스터로 모이거나 또는 떼어내야 하는 바로 그 기준점을 찾아내는 일이다. 예를 들어, 어떤 특정한 점의 집단이 있다고 할 때 그 점들 사이에서는 속성상 서로 독립된 관계를 유지하고 있고 그 점들을 포함한 모집단 내의 다른점들과는 서로 깊은 상관관계를 유지하는 점들이 있다면 그야말로 그런 점들은 서로 다른 클러스터를 형성하도록 분리시킬 수 있는 것이다. 즉, 다시 말하면, 그런 점들이야말로 각각의 클러스터 내에서의 센트로이드(key point)가 될 수 있다. 여기에서 key-cluster[32] 분석법을 수정 소개해 보면 다음과 같은 기본적 목적을 갖고 있음을 알 수 있다.

1) 모집단으로 각 클러스터를 특징지울 상호 대등한 위치에 놓인(mutually collinear) 점들을 선정하기 위함.
2) 모집단 내의 모든 점들과 상호 상관관계를 유지할 수 있으며 모든 점들의 통일성(communality)을 제시할 수 있는 최소한의 점을 찾기 위함(가장 소수의 지배집단)
3) 클러스터의 속성을 최소로 만족시킬수 있는 점들 이상으로 구성된 집단을 선정하기 위함.

key-cluster 분석의 목적을 만족시킬 그런 key-cluster를 찾기 위한 기법으로 커어널 기법을 소개한다. 커어널 개념은 본노이만(J. von Neumann)과 모오겐스턴(O. Morgenstern)[39]에 의해 게임이론에서 최초로 소개되었다. 그들은 커어널 개념을 게임을 풀 수 있는 가능한 해결책을 정의하기 위해서 사용하였다. 게임에 있어서 가능한 결과를 표현할 수 있는

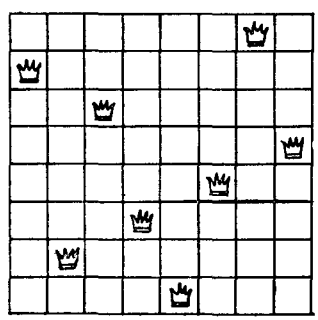
방향성을 갖는 그래프와 그 그래프를 형성하고 있는 점들이 있다 할 때, 점 X와 점 Y는 게임을 수행하는 사람이 Y라는 방법 보다는 X라는 방법을 더 선호 하고 실지로 Y보다 X가 더 효과적일 때 선으로 연결되어 진다면, 우리는 다음과 같은 특성을 갖는 집단을 찾아낼 수 있을 것이다 즉, 같은 집단 내에서는 어느 방법도 다른 어떤 방법보다 더 우수한 방법이 존재하지 않으며(독립성), 집단 밖의 어느 방법 보다도 집단 내의 방법이 우선 선호성을 갖고 있다. (지배성).

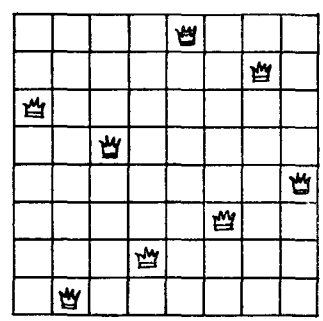
위에서 보여진 예처럼 커어널의 속성은 독립성(independence)과 지배성(*domination*)으로 정의될 수 있다.[3]

$G=(X, \Gamma)$로 정의된 그래프에서 X의 부분집합인 S는 집합 S내의 어느 두 점도 인접되지 않을 때 그래프 G의 독립집합이라 한다.

$$\Gamma(S) \cap S=\phi$$

독립집합 S는 S보다 더 큰 독립집단 $S'(S'>S)$가 존재하지 않을 때 가장 큰 집합이 된다. '가우스(Gauss)의 '8개의 여왕'[3]의 예를 들 수 있다. 문제는 '체스(chess) 게임판 위에 8개의 여왕이 있고 이들을 서로 잡아먹지 못할 위치에 배열할 수 있는가?'이다. 이 문제는 다음과 같은 의미를 갖는다. 즉, 64개의 점으로 구성된 그래프에서 $y \in \Gamma(x)$란 X가 놓인 위치로 부터 같은 열, 행 또는 대각선 상에 놓인 모든 점 (y)의 집합이라 할 때 최대의 독립집합을 구하는 문제이다.





독립집합의 개념은 재고창고 문제 해결에도 응용될 수 있다.[4] n개의 화학품$c_1, c_2, \ldots\ldots c_n$을 생산하는 공장이 있다고 할 때, 화학품 중 일부가 가까이 놓이거나 접촉할 경우 폭발하거나 위험한 상태가 된다면 우리는 위에 언급한 독립성의 개념대로 그라프 G의 집합으로 $\{v_1, v_2, \ldots\ldots v_n\}$을 만들 수 있고 화학품 c_i와 c_j가 서로 같이 놓이지 못할 경우 각 화학품을 대변할 점 v_i와 v_j가 인접한 점으로 하면 문제를 풀 수가 있다.

커어널의 속성 중 다른 하나인 지배성에 대해 살펴보면 그래프 $G=(X, \Gamma)$에서 부분집합 S가 다음과 같이 정의될 때 집합 S는 S에 속하지 않은 어떤 점 x에 대해서도 지배성이 있다고 말할 수 있다.

$$\Gamma(x) \cap S \neq \phi$$

지배집합에 대한 문제도 여러곳에서 나타난다. 전략지역에 여러개의 레이더 기지를 설치할 때 이 원리가 적용된다.[3] 가능한한 적은 수의 레이더 기지로 전체를 관측하고자 할 때, 레이더를 설치할 수 있는 모든 지점을 집합 G의 점으로 표시하고 서로 시야에 들어오는 두 레이더 기지 X와 Y를 한개의 선으로 이어놓는다면 어느 두 지점도 서로 관측되지 않는 그런 점들로된 집합을 만들어 낼 수 있다.

이와 비슷한 상황이 위원회 임원을 선출할 때도 나타난다.[13] 각 회원들로 부터 자신의 의견과 필요성을 대변할 임원을 한명씩 선정하게 하고(그래프 내에서 두 점을 선으며 이어준다.) 이렇게 하여 구해진 그래프 속에서 그 집단을 대표할 가장 적은 수의 지배집합을 선출하면 되는 것이다.

마지막으로 동시에 독립성과 지배성을 만족하는 일련의 집합이 있을 때 이것을 커어널 즉, 그래프의 안정된 요소들의 집합이라 부른다. 이런 커어널의 개념은 게임이론에서 많이 쓰이고 있으나 자동화된 도서관 시스템에서 활용된 예가 있다.[17] 도서관 내에 있는 모든 책을 하나의 점으로 하는 그래프를 만들고 임의의 두 점 x, y가 서로 인용관계에 있을 때 선으로 이어준다면 커어널 집합에 포함된 최소의 책으로 부터 도서관 전부에 접근할 수 있게 만들 수 있는 것이다.

다음 장에서는 이러한 속성을 갖는 커어널 기법을 실제로 정보검색 시스템에 활용한 예를 설명하면서 정보검색 시스템 설계의 새로운 가능성을 제시해본다.

Ⅳ. 커어널의 정보검색 응용

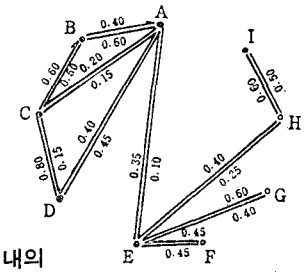
커어널 기법을 활용한 검색 시스템의 개발은 다음과 같은 가설로 시작될 수 있다.

가설 : 데이터베이스 내의 문헌들은 문헌간의 상관계수에 따라 여러개의 클러스터로 나뉘어지며 각 클러스터는 그 클러스터의 속성을 특징지을 커어널 집합으로 대변되어진다. 뿐만 아니라 문헌상관계수의 기준치를 낮추어 줌에 따라 커어널 집합은 새로운 클러스터를 형성하고 또 다른 상위커어널 집합을 생산하여 결국 데이터베이스 내의 문헌을 계층적으로 구축하여 문헌 탐색지도를 만들 수 있다.[14]

데이터베이스 내의 문헌을 여러개의 클러스터로 나누기 위해선 데이터베이스내의 문헌들을 그들의 상관관계에 따라 문헌 행렬도를 만들어야 한다. 상관관계계수를 측정하는 공식들 중, 비대칭 상관계수를 사용함으로써 문헌의 상·하 개념 및 포함 관계를 좀 더 명확히 규정지을 수 있다.

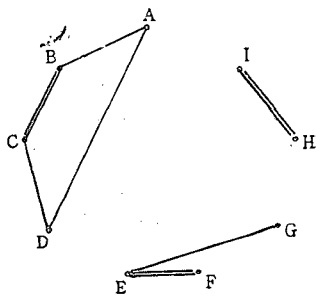
$$P_{ij}=\frac{|d_i \cap d_j|}{|d_i|}$$

문헌간의 상관계수에 의한 방향성은 상관계수 기준치의 상, 하향 조정에 따라 변할 수 있으며, 본 커어널 기법의 특징은 기준치 0에서 구해진 커어널 집합을 구분지을 수 있는 선에서 자동으로 기준치를 구할 수 있는 장점을 갖는다. 다음은 그 과정을 그래프를 통하여 설명함으로써 보다 높은 이해를 구할 수 있다.



커어널 집합 내의
요소를 분리할 수
있는 상관계수
기준치(threshold)=
0.425

커어널 집합=
{C, E, I}



다음의 프로그램은 연결된 그래프 내에서 커어널 집합을 찾아주는 프로그램이다.

Table 1 PROGRAM FOR SEARCHING KERNELS

```
PROGRAM KERNEL (INPUT, OUTPUT) ;
TYPE
   ST=1.. MAXINT ;
   kERTEX=ARRAY[ST, ST] OF INTEGER ;
   NEIGHBOR=SET OF ST ;
   LIST= ^NEIGHBORSET ;
   NEIGHBORSET=RECORD
                        INDEX : INTEGER ;
                        INFO : NEIGHBOR ;
                        LINK : LIST
                      END ;
   POINT=^KERNELS ;
   KERNELS=RECORD
```

```
                INFO : NEIGHBOR ;
                LINK : POINT
            END ;
VAR
   I, J, K, N : INTEGER ;
   VER : VERTEX ;
   TOTAL, COMP, NEIG : NEIGHBOR ;
   NEIST : NEIGHBORSET ;
   LLAST, LFIRST, LSTART : LIST ;
   PLAST, PFRONT, PFIRST, PSTART : POINT ;
   KER : KERNELS ;
   STATE : BOOLEAN ;
PROCEDURE INSERT ;
BEGIN
   NEW(PLAST) ;
   IF NEIG=[ ] THEN
      PLAST^.INFO : =[I]
   ELSE
   PLAST^.INFO : TOTAL-NEIG ;
   PLAST^.LINK : =NIL ;
   IF PSTART : NIL THEN
      BEGIN
         PSTART : =PLAST ;
         PFIRST : =PLAST
      END
   ELSE
      BEGIN
         PFIRST^.LINK : =PLAST ;
         PFIRST : =PLAST
      END
   END ;
BEGIN
   WRITE('THE NUMBER OF VERTICES IN A GRAPH : ') ;
   READ(TTY, N) ;
      FOR I : =1 TO N DO
```

```
    BEGIN
      FOR J: =1 TO N DO
        BEGIN
          READ (K);
          VER[I, J]: =K
        END;
      READLN
  END;
TOTAL: =[ ];
FOR I : =1 TO N DO
  TOTAL: =TOTAL+[I];
LSTART: =NIL;
PSTART: =NIL;
FOR I : 1 TO N DO
  BEGIN
    NEIG: =[ ];
    NEW(LLAST);
    LLAST^.INDEX: = I;
    LLAST^.LINK: =NIL;
    IF LSTART=NIL THEN
      BEGIN
        LSTART: =LLAST;
        LFIRST: =LLAST
      END
    ELSE
      BEGIN
        LFIRST^.LINK: =LLAST;
        LFIRST: =LLAST
      END;
    FORJ: =1 TO N DO
      IF VER[I, J]=1 THEN
        NEIG: =NEIG+[J];
    LLAST^.INFO: =NEIG;
    INSERT
  END;
```

```
PFIRST : =PSTART ;
WHILE PFIRST<>NIL DO
   BEGIN
      STATE : =FALSE ;
      LFIRST : =LSTART ;
      WHILE (LFIRST<>NIL) AND NOT STATE DO
         BEGIN
            COMP : =(LFIRST^. INFO) AND (PFIRST^. INFO) ;
            K : =LFIRST^. INDEX ;

      IF K IN PFIRST^. INFO THEN
         BEGIN
            IF COMP <>[ ] THEN
               BEGIN
                  PFIRST^. INFO : =PFIRST^. INFO-COMP ;
                  LFIRST : =LSTART
               END
            ELSE
            LFIRST : =LFIRST^. LINK
               END
            ELSE
               BEGIN
                  IF COMP<>[ ] THEN
                     LFIRST : =LFIRST^. LINK
               ELSE
                  STATE : =TRUE
               END
            END ;
         IF STATE=TRUE THEN
            BEGIN
               IF PFIRST=PSTART THEN
                  BEGIN
                     PSTART : =PSTART^. LINK ;
                     PFIRST : PSTART
                  END
```

```
            ELSE
               BEGIN
                  PFIRST : =PFIRST^. LINK ;
                  PFRONT^. LINK : =PFIRST
               END
            END
         ELSE
            BEGIN
               PFRONT : =PFIRT ;
               PFIRT : =PFIRT^. LINK
            END
         END ;
      PFIRST : =PSTART ;
      WHILE PFIRST<>NIL DO
         BEGIN
            PLAST : =PFIRST^. LINK ;
            PFRONT : =PFIRST ;
            WHILE PLAST<>NIL DO
               IF PFIRST^. INFO=PLAST^. INFO THEN
                  BEGIN
                     PLAST : =PLAST^. LINK ;
                     PFRONT^. LINK : =PLAST
                  END
         ELSE
            BEGIN
               PLAST : =PLAST^. LINK ;
               PFRONT : PFRONT^I. LNK
            END ;
         PFIRST : =PFIRST^. LINK
      END ;
   IF PSTART=NIL THEN
      WRITE('THERE ARE NO KERNELS IN THIS
      GRAPH') ;
      K : = 0 ;
      PFIRST : =PSTART ;
```

```
WHILE PFIRST<>NIL DO
  BEGIN
    K : =K+1 ;
    WRITE('KERNEL', K : 2, ' =<') ;
    FOR I : 1.TO N DO
    IF I IN PFIRST^.INFO THEN
      WRITE(I : 2) ;
      WRITELN('>') ;
      PFIRST : =PFIRST^.LINK
    END
  END.
```

위와같은 방법으로 구해진 커어널 집합과 각각의 클러스터를 기본으로 하는 상위 클러스터를 구축하기 위해 기준치를 다시 낮추고 낮추어진 기준치에 따라 하나로 연결된 커어널들만의 집합으로 새로운 커어널 집합(super kernel)[14]을 생산 새로운 클러스터를 형성, 하위 클러스터와의 연결 지도를 형성한다. 이런 방법을 계속 반복하여 결국 전체 데이터베이스를 재구성, 문헌 탐색 지도를 만든다. (다음의 지도 구성 규칙과 그에 의해 구축된 탐색 지도를 참조할 것)

실제로 1331개의 문헌으로 구성된 식이요법 관련 문헌 중 102개의 인용율이 높은 문헌을 기본 데이터베이스로 구축, 위의 방법을 통하여 탐색 지도를 만들어 보고 부울 검색과 비교한 검색효율과 결과를 측정해 보았다.

MAPPING PROCEDURE

STEP 0 : If there are no vertices to group or any vertices cannot be clustered together on the lowest threshold value then goto exit, else continue ;

STEP 1 : i : = 0 ;

STEP 2 : Find maximum threshold value not to change the graph state (connected) ;

STEP 3 : Find the appropriate kernel set (arbitrarily choose one or more) ;

STEP 4 : If i= 0 then goto stop 8, else continue ;

STEP 5 : Select the elements of kernel sets from each connected graphs ;

STEP 6 : Reform new graphs with selected vertices ;

STEP 7 : Goto stop 0 ;

STEP 8 : Recluster (raise the threshold value) to group vertices in the way all elements in kernel set are separated ;

STEP 9 : i : =i+1 ;

STEP 10 : Goto step 2 ;

EXIT.

THE RATIO OF PRECISION AND RECALL IN SEARCHING THROUGH MAP

Section	Question #1		Question #2		Question #3	
A, B, C	Precision	Recall	Precision	Recall	Precision	Recall
A+B+C	.625	.714	.692	.600	.714	.526
A+B	.500	.429	.556	.333	.667	.316
B+C	.800	.571	.889	.533	.900	.474

1. Section : a sequence of relevant documents to a query is divided by three sections

   Section A : set of documents which contain broad concept to a query

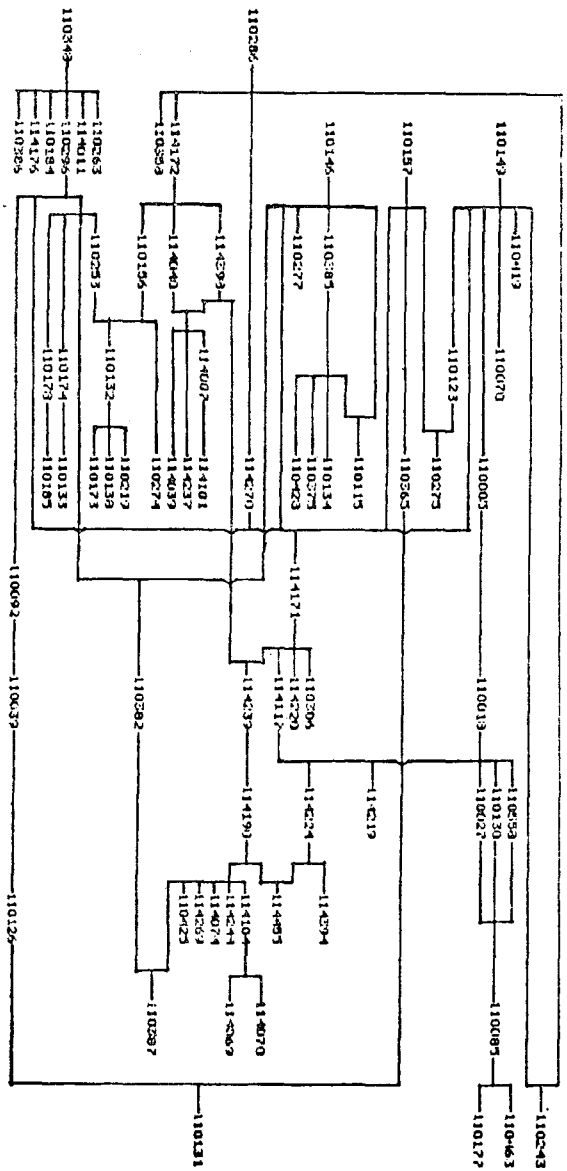
   Section B : set of documents which contain the most adequate concept to a query

   Section C : set of documents which are rather specific to a query

2. Questions : these questions are provided by Norwich-Eaton Pharmacautical, Inc.

   Question #1 : "Elemental Diets and Caloric Need in Patients with Head In Juries"

   Question #2 : "Food Allergies and Vivonex or Other Elemental Diets"



문헌탐색지도

Question #3 : "The Use of Vivonex or Other Elemental Diets in Newborn Infants"

탐색 지도에 의해 검색된 문헌의 집합은 부울 대수의 단점으로 지적된 상관도에 따른 배열을 얻을 수 있었으며 검색의 레벨을 조정함으로써 재현율과 정확율을 조정, 효율을 높일 수 있었다.

TEST OF RETRIEVAL

Question #1 : "Elemental Diets and Caloric Need in Patients with Head Injuries"

Total size of documents : 102 documents

Used terms by code :

1223, 1290, 1293, 1406, 1430, 1578 and 1759

SEARCH BY BOOLEAN

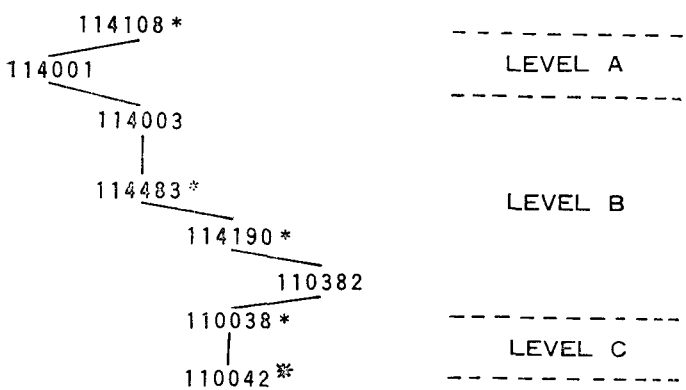
Combination :

1406 OR (1223 AND 1290) OR [1430 AND 1578 AND (1293 OR 1759)]

Result by code :

110038, 110042, 114051, 114065, 114108, 114190 and 114483

SEARCH BY PROPOSED



relevant documents.

Ⅴ. 결언

정보검색 시스팀에 있어서 부울대수의 단점을 보완하고 검색의 효율을 높이기위해 클러스터링 기법이 소개되었으며 클러스터링 기법의 문제점과 계층적 구조의 장점을 최대한 살린 커어널 기법을 정보검색 시스팀에 응용하여 보았다. 그래프 이론과 게임 이론에 적용되어온 커어널 기법을 그 속성에 의해 계층적 커어널로 확장하여 검색 시스팀에 활용하여 본 결과 클러스터링 기법이 대규모의 데이터베이스에 활용하기 힘들다는 단점에도 불구하고 검색의 효율을 높일 수 있었으며, 주관적 개념인 적합성의 객관화를 위한 다층 레벨에 의한 검색으로 이용자에 따라 높은 정확율, 높은 재현율을 선택 탐색할 수 있었다.

뿐만 아니라 문헌 탐색 지도를 마련함으로써 일단 구축된 지도를 문헌 탐색 지도에서 의미 지도로의 전환을 꾀할 수 있었다. 결국 기존의 계층적 탐색에 보다 효율을 높인 다층 레벨 구조로 바꿈으로써 탐색 시간과 효율을 동시에 높일 수 있었다.

앞으로 이와 유사한 연구가 계속 되길 바라며 문헌을 대상으로한 분석보다는 디소오러스를 다층 레벨의 의미 지도로 변환시킬 수 있는 기법으로 활용되기를 바란다.

참 고 문 헌

1. Belnap, N.P. "An analysis of question : preliminary report." *Scientific Report TM-1287, SDC* (Santa Monica, 1963).
2. Belnap, N.P. and Steel, T.B. *The Logic of Questions and Answers.* (New Haven : Yale University, 1976).
3. Berge, Claude. *Graphs and Hypergraphs.* [translated by E. Minieka] (New York : North-Holland, 1976).
4. Bondy, J.A. and Murty, U.S.R. *Graph Theory with Applications.* (New York : North-Holland, 1976).

5. Can, Fazil and Ozkarahan, Esen A. "Two partitining type clustering algorithms." *JASIS* 35(1984) : 268~276.
6. Cooper, D.B. *A Geometrical Model for Information Retrieval.* (Ph. D. dissertation, Case Western Reserve University, 1973).
7. Doyle, L.B. "Semantic road maps for literature searches." *Journal of the ACM* 8(1961) :
8. Garfield, E. "ISI is the studying the sturcture of science through co-citation analysis." *Current Contents* 7(1974) : 5~10.
9. Gerson M. "Cliqueing-a technique for producing maximally connected clusters." *JASIS* 29(1978) : 125~129.
10. Goffman, Willim." Indirect method of information retrieval." *Information Storage and Retrieval* 4(1968) : 361~373.
11. Goffman, Willaim, Verhoeff, J. and Belzer, Jack. "Inefficiency of the use of Boolean functions for infomation retrieval systems." *Communications of the ACM* 4(1961) : 557~559.
12. Goffman, William and Warren, K.S. *Scientific Information Systems and the Principle of Selectivity.* (New York : Praeger, 1980).
13. Harary, F. etal. *Structural Models : an Introduction to the Theroy of Directed Graphs.* (New York : John wiley, 1965).
14. Jeong, Jun Min. *The Ecology of the Scientific Literature and Information Retrieval.* (Ph. D. dissertation, Case Western Reserve University, 1985).
15. Kessler, M.M. "Bibliographic coupling between scientific papers." *American Documentation* 14(1963) : 10~25.
16. Litofsky, B. *Utility of Automatic Classification Systems for Information Storage and Retrieval.* (Ph. D. Dissertation University of Pennsylvania, 1969).
17. Liu, C.L. *Introduction to Combinational Mathmatics.* (New York : Mceraw-Hill, 1968).
18. Maron, M.E. and Kuhn, J.L. "On relevance, probabilistic indexing and information retrieval" *Journal of the ACM* 7(1960) : 216~244.
19. Miyamoto, S. and Nakayama, K. "A technique of two-stageclustering applied to environmen and related methods of citation analysis." *JASIS* 34(1983) : 192~201.
20. Montgomery, C.A. "Linguistics and information science." *JASIS* 23 (1972) : 195~219.

21. Noreault, T. etal. "Automatic ranked output from Boolean searches in SIRF." *JASIS* 28(1977) : 333~341.
22. Norma, Elliot J. *Centroid Scaling of Citation Data.* (Ph. D. dissertation, University of Michigan,1982).
23. Salton, Gerard. "Progress in automatic informtion retrieval." *IEEE Spectrum* (1965) : 90~103.
24. Salton, Gerard and McGill, M.J. *Intoduction to Modern Information Retrieval.* (New York : McGraw-Hill, 1983).
25. Shaw, William M., gr. personal communication.
이것은 노드(문헌)와 링크(상관계수, 기준치)의 갯수의 상관관계에 따라 최소 N개의 링크가 n개의 노드를 위해 존재해야 한다는 최소 N을 만족시키는 기준치 값을 찾아내는 기법을 말한다.
26. Small, Henry. "Co-citation in the scientific literature : a new measure of the relationship between two documents." *JASIS* 24(1973) : 265~269.
27. Small, Henry and Griffith, B.C. "The structure, I. identifying and graphing specialties." *Science Studies* 4(1974) : 17~40.
28. Smith, Linda C. "Artificial intelligen in in formation retrieval systems." *Information Processing and Management* 12(1976) : 189~222.
29. Sparck Jones, Karen. "Collection properties influencing automatic term classification performance." *Information storage and Retrieval* 9(1973) : 510~513.
30. Sparck Jones, Karen. "Some throughts on classification for retrieval." *Journal of Documentation* 26(1970) : 89~101.
31. Sparck Jones, Karren and Kay, M. *Linguistics and Information Service.* (New York : Academic Press, 1973).
32. Tryon, Robert C. and Bailey, Daniel E. *Cluster Analysis.* (New York : McGraw-Hill 1970).
33. Van Rijsbergen, C.J. "Automatic classification in formation retrieval." *Drexel Library Quarterly* 14(1978) : 75~89.
34. Van Rijsbergen, C.J. *Automatic Information structuring and Retrieval.* (Ph. D. dissertation University of Cambridge, 1972).
35. Van Rijsbergen, C.J. *Information Retrieval,* 2nd ed. (London : Butterworth, 1979.
36. Van Rijsbergen, C.J. and Jardine, N. "The use of hierarchic cluster-

ing in information retrieval." *Information and Retrieval* 7(1971) : 217~240.

37. Van Rijsbergen, C.J. and Sparck Jones, K. "A test for the sparation of relevant and non-relevant documents in experimental retrieval collections." *Journal of Documentation* 29(1973) : 251~257.
38. Van Vyzin, J. *Classification and Clustering*. (New York : Academic Press, 1977).
39. Ron Neumann, J. and Morgenstərn, O. *Theory of Games and Economic Behavior*. (Princenton : Princeton University Press, 1944).
40. Yu, C.T. "A clustering algorithm based on user queries." *JASIS* 25(1974) : 218~226.

The Study of Kernels in Information Retrieval

Jun Min Jeon*

ABSTRACT

The kernel technique in game theory is introduced and modified in the notion of super kernel, which creates searching maps through the hierarchic clustering technique. The results show more improved retrieval efficiency in terms of precision and recall.

* Assit. professor, Chonnam National University.