Research of Industrial Technology
Vol. 22, No. 1, 2004. 6

한국 컴퓨터정보학회 논문지
8/2, 2003. 6

연관 규칙 탐사 기법을 이용한 해양 전문 검색 엔진에서의 질의어 처리에 관한 연구

하 창 승*, 윤 병 수**, 류 길 수***

A Research on User's Query Processing in Search Engine for Ocean using the Association Rules

Ha, Chang Seung * Youn, Byung Soo ** Rhyu, Keel Soo ***

요 약

최근 여러 가지 정보들이 WWW를 경유하여 제공되고 있기 때문에 검색엔진의 필요성은 점점 커지고 있다. 그러나 대부분의 검색엔진은 정보의 추출을 위해 웹 문서와 사용자 질의를 단순 패턴비교 방법을 사용함으로써 검색엔진의 효율은 비교적 낮은 편이다. 일반적으로 사용자의 검색 목적에 따라 다른 검색엔진이 사용되기 때문에 여러 전문검색엔진을 개발하고 있지만 대부분의 검색엔진들이 사용자의 요구를 제대로 반영하고 있지 못하다. 본 연구에서는 웹 데이터마이닝의 연관규칙을 이용하여 사용자 질의를 처리하는 해양전문검색엔진을 제안한다. 데이터 마이닝 분야에서 주로 연구되어온 연관규칙탐사 기법은 지지도와 신뢰도에 따라 연관자료의 확신도를 측정할 수 있기 때문에 웹 문서 사이의 관련성을 입증하는데 이 규칙을 적용하여 기존의 검색 방법에서 자료의 재현률과 정확률을 개선하였다.

Abstract

Recently various of information suppliers provide information via WWW so the necessary of search engine grows larger. However the efficiency of most search engines is low comparatively because of using simple pattern match technique between user's query and web document. A specialized search engine returns the specialized information depend on each user's search goal. It is trend to develop specialized search engines in many countries. However, most such engines don't satisfy the user's needs. This paper proposes the specialized search engine for ocean information that uses user's query related with ocean and the association rules in web data mining can prove relation between web documents. So this search engine improved the recall of data and the precision in existent search method.

* 동명대학 정보통신계열 조교수, ** 한국해양대학교 대학원 컴퓨터공학과 박사과정
*** 한국해양대학교 기계정보공학부 교수

I. 서론

최근 기업, 단체 및 개인에 이르기까지 많은 정보 제공자들이 인터넷을 통하여 정보를 제공함에 따라 웹 상의 정보 양이 급속히 증가하고 있고 웹을 통한 정보의 검색 또한 점점 일반화되고 있다. 정보량의 급속한 증가는 사용자에게 필요한 정보 검색의 어려움과 다량의 정보로부터 핵심 지식의 창출 및 개인화 된 정보 제공의 문제를 야기 시키고 있다. 이에 대해 기존의 상용 검색엔진들은 대부분 방대한 정보의 양을 가진 인터넷에서 사용자들이 필요로 하는 정보를 제공하기 위해 주어진 질의어와 웹 상의 문서간의 패턴 비교를 통하여 일치하는 정보를 검색하는 기법을 사용함으로써 검색 효율이 비교적 낮다[1]. 특히 오늘날 정보 요구자 자신과 관련된 특정 영역에 있어 사용되는 전문용어 혹은 이와 유사언어에 대한 검색에 대하여 기존의 검색엔진들은 연관성 높은 검색결과를 제대로 지원하지 못하고 있다[2].

전문검색엔진은 검색 목적에 따라 영역별 전문지식을 검색결과로 제공할 수 있어야 한다. 분야별 전문검색엔진의 다양화는 외국에서도 이미 대중화된 인터넷 서비스 중 하나로 미국만 해도 1,800여 개의 전문검색엔진이 사용되고 있다. 예를 들어 뉴스의 헤드라인만 검색해 주는 사이트(www.moreover.com), 연방법과 정부의 웹사이트만을 전문적으로 검색해 주는 사이트(www.findlaw.com), 과학기술과 관련된 정보를 제공하는 사이트(www.biolinks.com) 등으로 특성화되어 가고 있는 추세이다[3].

하지만 대부분의 전문검색엔진에서 제공하는 정보의 수준은 전문지식을 필요로 하는 사용자의 기대보다 미흡한 경우가 많다. 이는 검색엔진이 해당 질의어에 대해 로컬 데이터베이스를 통해 검색함으로써 한정된 정보만을 제공하거나 질의어와 연계된 관련성 있는 정보를 함께 제공하지 못하는 문제점을 나타내고 있다[4]. 이에 본 연구에서는 사용자들이 해양관련 질의어에 대하여 데이터마이닝 기법 중 연관규칙탐사기법을 이용하여 사용자들의 질의어에 대하여 해당분야에서 그 질의어와 관련성 정도가 높은 유사질의어의 정보를 추가로 제공함으로써 사용자 질의어에 대하여 보다 높은 재현률과 관련성 깊은 정보를 제공하고자 한다.

II. KDW의 개념 및 기법

최근 학계, 산업계에서는 KDW(Knowledge Decovery in Web)를 통한 지식획득에 대한 연구가 꾸준히 진행되어 왔다. KDW는 웹으로부터 유용한 정보 및 지식을 발견하기 위해 데이터 선택 및 정제, 보완, 변환, 마이닝 기법의 적용, 모형의 평가라는 과정을 거쳐 지식 발견 프로세스를 통하여 이루어진다[5]. 웹 데이터마이닝은 로컬 데이터베이스가 아닌 웹 상에서 대량의 데이터에 내포되어 있는 데이터간의 의미 있는 상호관련성과 유용한 패턴을 추출한다.

1. KDW 관련 연구

웹 사용자의 성향파악을 통한 개인화된 정보 제공과 웹 에이전트의 학습을 위해 웹 데이터마이닝에서 이용하는 알고리즘은 연관 규칙(association rule), 군집화(clustering), 순차패턴(sequential pattern), 웹 방문패턴(web navigation patterns) 등이 있다[6].

먼저 군집화란 유사한 특성을 갖는 여러 객체를 몇 개의 그룹으로 클러스터링 하는 마이닝 기법으로 비교사학습(unsupervised learning)의 특징을 지니면서 정해지지 않은 다량의 데이터 집합에 대해 데이터베이스의 제한적인 요소들의 충족 정도가 평가 기준이 된다. 웹 마이닝에서 군집화의 대상 객체는 웹 사이트를 방문한 사용자, 웹 사이트를 구성하는 페이지 등이 되며 웹 검색엔진에서 카테고리 생성에 이용된다. 군집화 알고리즘은 크게 파티션 최적 알고리즘과 계층적 클러스터링 최적 알고리즘으로 나뉘어진다. 파티션 최적 알고리즘은 K개의 모든 가능한 파티션을 열거한 후 군집화가 얼마나 잘 형성되었는지 나타내는 척도로서 함수값이 가장 좋은 것들로 그룹들을 정한다. 대표적인 파티션 알고리즘은 숫자 속성데이터를 군집화하는데 가장 오래 이용된 K-평균 군집화 알고리즘이었다. K-평균 군집법은 n개의 속성으로 구성된 각각의 레코드를 벡터로 표시하며 n차원의 공간에 나타내어 유사한 특성을

가진 레코드들을 서로 근접시킨다. 먼저 각 개체를 가장 가까운 중심에 할당하는 방법으로서 K개의 초기 중심점을 선택한다. 그리고 각 개체를 가장 가까운 중심점을 갖는 군집으로 할당한 후 새로운 군집의 중심점을 정의한다. 이 때 각 개체의 할당에 변화가 없을 시 전 단계를 통해 최종적으로 K개의 군집을 형성한다. 또한 계층적 클러스터링 최적 알고리즘은 하향식과 상향식 알고리즘이 있는데 상향식 알고리즘은 우선 모든 n개의 데이터가 n개의 서로 다른 그룹이라 가정한 후 그룹간의 유사성을 보고 가장 유사한 두 개의 그룹을 합병하여 그룹 수를 줄여 가는 방법이다. 유사성을 평가하는 함수는 두 개체간의 거리로서 나타내는데 거리 결정은 유클리드 거리, 맨해튼 거리, 클래스간 평균거리, 클래스 내 평균거리 측도 방법이 있으며 계층적 클러스터링 최적 알고리즘의 군집방법에는 최단연결법, 최장연결법, 평균연결법, 중심연결법, 중위수연결법 등이 있다[7].

순차패턴은 각 사용자들의 한 트랜잭션 안에서 발생하는 페이지간의 연관규칙에서 시간적 변이 개념이 추가된 것이다. 즉 연관규칙은 트랜잭션 안에서 어떤 페이지들간의 상호 연관적 관계를 통해 관련성을 평가하는 반면 순차패턴은 트랜잭션 상호간의 관계를 평가하는 것이다. 각 사용자들의 순차적인 트랜잭션을 사용자 순차집합이라고 하고 순차가 특정 고객에 대한 사용자 순차 집합에 속해 있다면 그 사용자는 시 순차를 지지한다고 한다. 순차에 대한 지지도의 정의는 순차를 지지하는 전체 사용자들의 수이고 지지도를 만족하는 순차를 빈발순차라 한다. 또한 순차 패턴 탐색은 사용자가 정의한 최소 지지도를 만족하는 모든 순차들 사이에서 최대 순차들을 순차패턴이라 하고 순차패턴은 과거 사용자가 접속한 세션 정보를 통해 추출할 수 있다.

웹 방문패턴은 웹사이트에 존재하는 문서 페이지들의 구성 경로를 찾는 방법을 제공해 준다. 이는 앞의 연관 규칙과 유사한 특성을 보이지만 페이지간의 순차적인 관계가 있다는 점만 다르다. 또한 웹 방문 패턴에서 사용되는 최소지지도라는 변수 값은 사용자 세션에서 얻은 패스 중에서 패턴으로서 의미가 있는 최소 발생 빈도 수를 의미한다.

2. 연관규칙 탐사 기법

연관규칙 탐사 기법은 항목집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 찾는 방법으로써 데이터 혹은 서비스간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고자 할 때 이용되는 기법이다. 연관규칙 탐사 기법에서는 데이터나 서비스의 트랜잭션 로그(log)로부터 데이터간의 연관성 정도를 측정하여 사용자의 요구에 대해 연관성이 높은 추가적인 요구들을 그룹화 하여 동시에 제공시킬 수 있도록 데이터 상호간에 연관성을 부여하고 있다[8].

① 가설(hypothesis)
- ItemSet I의 부분집합 X에 대해 $X \subseteq T$이면 X를 만족한다고 정의한다.
- Itemset X ($X \subseteq I$)를 만족시키는 D의 트랜잭션 수를 $|X|$로 표기한다.
- X, Y $\subseteq$ I에 대한 R은 $X \cap Y = \varnothing$의 특성을 갖는다.

② 측정(Measure)
- Support (연관규칙 X⇒Y에 대한 지지도)

$$S = \frac{|X \cup Y|}{N}$$

- Confidence(연관규칙 X⇒Y에 대한 신뢰도)

$$C = \frac{|X \cup Y|}{|X|}$$

연관규칙탐사기법은 그림 1과 같이 기본적으로 2단계로 구성된다. 1단계에서는 사용자가 미리 정의한 최소 지지도를 만족하는 데이터 항목조합들만 추출하고 2단계에서는 1단계에서 얻은 데이터의 부분 집합에서 생성된 규칙 중 사용자가 정의한 최소 신뢰도를 만족하는 규칙들을 탐색하여 최종 규칙으로 정하게 된다. 연관규칙의 탐사기법의 성능은 1단계에서 결정되며 1단계에서 추출한 빈발 항목집합(large item sets)을 확인한 후에 연관규칙의 신뢰도는 2단계에서 평가된다. 예를 들어 $I=\{i_1, i_2, \ldots, i_k\}$를 항목집합이라고 하고 트랜잭션들로 이루어진 데이터 셋(data set) D가 존재할 때 각 트랜잭션은 고유한 트랜잭션 번호(TID)가 부여된다. X⇒Y 형식에서 $X \subset I$, $Y \subset I$이고 $X \cap Y = \varnothing$일 때 연관규칙은 지지도와 신뢰도를 바탕으로 트랜잭션 데이터 셋에서 각 항목간의 연관성을 찾는 것을 의미한다. 지지도는 전체 트랜잭션에 대한 X와 Y를 포함하는 트랜잭션 비율을 의미하고 신뢰도는 X를 포함하는 트랜잭션에 대한 Y를 포함하는 트랜잭션을 말한다. 사용자가 정한 최소지지도를 만족하는 항목집합을 빈발 항목집합이라 하고 최소신뢰도를 만족하는 빈발항목 집합이 있으면 유효한 연관규칙이 있다고 말한다.

III. 질의어 처리 알고리즘의 구현

1. 연관규칙을 이용한 추론엔진의 구성

연관규칙의 관련도를 결정짓기 위해 가장 많이 사용되는 알고리즘으로 Apriori 알고리즘이 있다. 이 알고리즘은 기본적으로 미리 사용자가 정의한 최소지지도 이상의 트랜잭션 지지도를 갖는 빈발항목 집합을 결정하고 이 집합 중에서 빈발항목 요소 상호간에 규칙성을 찾아내어 신뢰도를 생성하는 기법으로 요소 상호간의 관련 정도가 집합 상호간의 관련도를 결정하는 평가함수가 된다. Apriori 알고리즘에서 사용하는 중요한 법칙은 빈도수가 높은 항목의 집합의 모든 부분 집합도 빈도수가 높다는 사실이다. 만약 주어진 요소수가 n개가 있을 때 이 항목을 이용해 만들 수 있는 부분집합의 수는 2^n이다. 예를 들어 {a, b, c}의 모든 부분집합은 {}, {a}, {b}, {c}, {a, b}, {a, c}, {b, c}, {a, b, c}이다. Apriori 알고리즘에서 지지도의 계산은 우선 요소의 개수가 하나인 항목집합의 빈도수를 계산하고 이 집합 중에서 지지도를 만족하고 요소 수가 두 개인 후보 항목 집합의 지지도를 결정하는 방법으로 요소의 수를 증가시켜 나간다. 그러므로 요소수가 k인 항목에서 지지도를 만족하는 집합에 대해서만 요소수가 k+1인 후보 항목 집합의 지지도를 결정하고 지지도를 미달하는 항목집합은 후보그룹에서 탈락시킴으로서 조합 가능한 부분집합의 수를 줄여나간다. 하지만 본 연구에서 제안하는 연관규칙 추론은 검색엔진의 특성상 주어진 요소수가 많아야 두 개 이상을 넘지 않는다는 제약이 있기 때문에 고려해야 할 차수의 수는 더욱 단순해진다. 즉 n개의 집합이 있다면 이 항목을 이용해 만들 수 있는 순서적 의미를 지닌 두 요소 항목 조합은 n(n-1)개이다. 예를 들면 질의어 요소 수가 n인 집합에서 순서적 의미를 갖는 두 요소 항목 조합은 (a, b), (a, c), (b, a), (b, c), (c, a), (c, b)로 단순화된다. 그림 2는 사용자가 입력한 항목에 빈도수와 항목조합에 따라 지지도를 계산하고 기 정의된 최소지지도(preset)에 따라 후보 항목 집합이 선정되는 과정을 보여주고 있다.

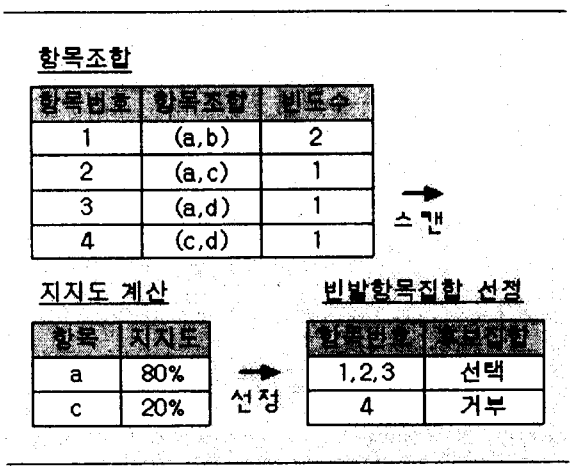

그림 2. 그룹화를 통한 항목조합 지지도 계산

Fig. 2 Calculate of support of item combination through grouping

초기 단계의 방대한 항목조합에서 빈발항목집합을 선정하기 위해 지지도 계산이 필요하며 이것은 다음과 같은 지지도 평가함수를 통해 결정된다.

$$\text{지지도}(Rb) = \sum_{i=1}^{n}\sum_{j=1} P(Si, Sj) \quad \because i \neq j \quad \cdots\cdots\cdots (1)$$

이것은 순서적 의미를 갖는 4가지 항목조합 트랜잭션에서 {a, b}항목은 {a} → {b} 규칙으로 표현되고 규칙의 왼편에 있는 항은 규칙의 오른편에 있는 항과 직·간접적으로 관련을 갖는다는 것이다. 이것은 전체 트랜잭션에 대한 항 {a}에 대한 항 {b}의 관련 확률로 나타낼 수 있음을 의미한다. 또한 항 {a}와 관련된 항 {b}, 항 {c}, 항 {d}의 지지도의 합은 전체 트랜잭션에 대한 항 {a}에 대한 지지도를 확률 값으로 나타나며 이 값은 최소 지지도인 임계값에 의해 후보 항목 집합으로 선정되거나 혹은 거부된다. 이때 식 (1)의 n은 각 트랜잭션의 항목조합에서 좌측 항을 포함하는 항목조합들의 개수이며 P는 각 항목조합들의 확률을 의미한다.

두 항목 상호간의 관련도의 정도를 결정하기 위해 수행할 두 번째 단계는 신뢰도의 계산이다. 신뢰도의 평가는 지지도를 만족하는 빈발항목집합 중에서 항목 요소 상호간의 관계연산이 AND연산인지 혹은 OR연산인지에 따라 다른 연관 가중치가 주어지며 연관 가중치를 갖는 신뢰도 평가함수를 통해 결정된다. 이 신뢰도의 결과 값은 기 정의된 임계값인 최소신뢰도에 따라 최종 유효 연관 집합으로 선택되거나 혹은 거부된다.

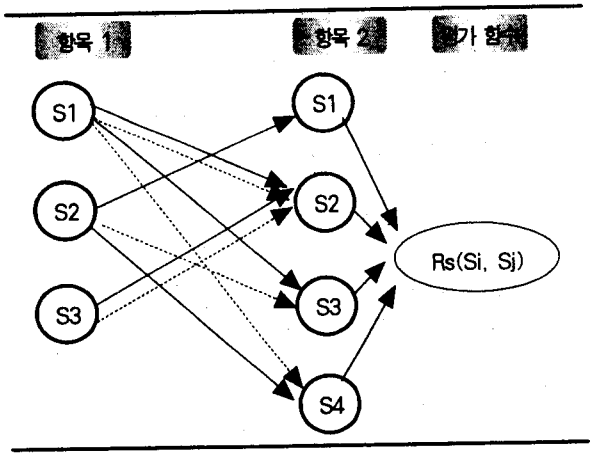

그림 3. 빈발항목 연산을 통한 신뢰도 계산

Fig. 3 Calculation of confidence rate through frequent item operation

〈그림 3〉은 임계값 이상의 지지도를 갖는 후보 집합에서 각 항목 요소 상호간의 관계를 나타낸다. 항목 요소간 관계 연산이 AND인 경우 실선으로, OR인 경우는 점선으로 표현하였다. 여기서 두 항목 요소간 관계 연산에 따라 다른 가중치를 주었으며 각 후보 집합의 신뢰도는 다음의 신뢰도 평가함수를 통해 결정된다.

$$\text{확신도}(Rs) = \left(\alpha * \frac{Si \rightarrow Sj\text{의 항목 수}}{Si\text{ 항목 수}} * An\right) + \left(\beta * \frac{Si \rightarrow Sj\text{의 항목 수}}{Si\text{ 항목 수}} * On\right) \cdots (2)$$

식 (2)에서 α는 AND 연산의 가중치, β는 OR 연산의 가중치를 의미하며, An은 AND 연산의 횟수 그리고 On은 OR 연산의 횟수를 의미한다. 후보 집합내 항목간의 신뢰도가 제시된 임계값 이상을 만족하는 경우에 두 항목간에 관련성이 있다고 정의할 수 있다. 표 1에서는 항목요소 상호간의 신뢰도와 유효 연관 집합의 선택여부를 나타내고 있다. 항목요소 조합 {S1, S2}와 {S3, S1}의 신뢰도는 임계값으로 주어진 최소신뢰도 30% 이상을 만족하여 관련성 있는 유효 항목집합으로 표시되고 있다. 표 1에 나타난 신뢰도는 AND 연산의 가중치로 1을 OR 연산의 가중치로 0.5를 부여한 경우이며 기호"√"는 선택, 기호 "X"는 거부를 의미하고 기호 "-"는 해당사항 없음을 의미한다.

요소2 / 요소1	S1		S2		S3		S4	
	신뢰도	선택	신뢰도	선택	신뢰도	선택	신뢰도	선택
S1	-	-	38%	√	25%	×	13%	×
S2	33%	√	-	-	17%	×	33%	√
S3	0%	×	75%	√	-	-	0%	×

표 4 그림 3을 이용한 신뢰도 평가 결과 예

Table. 1 Example of evaluation result using fig. 3

2. 검색엔진의 구성

본 연구에서 구현한 전문 검색엔진은 〈그림 4〉와 같이 세 부분으로 구성되어있다. 즉 인터넷상의 사이트 정보들을 추출하여 색인 데이터베이스로 재구성하는 로봇 에이전트 부분과 색인 데이터베이스를 이용하여 사용자의 검색요구를 처리해 주는 검색 에이전트 부분 그리고 검색된 정보를 사용자의 검색요구에 대한 부합 정도를 측정하는 연관 규칙탐사 기법 기반의 추론엔진 부분이다.

로봇 에이전트는 URL 데이터베이스로부터의 사이트 정보를 참조하여 해당 웹 서버들을 탐색하여 원시자료(Raw Data)를 수집한다. 그리고 수집된 자료는 색인 구축기(Index Builder)에 전달되어 색인 데이터베이스 구축에 이용된다.

검색 에이전트는 다시 프리젠테이션 계층과 트리거 계층으로 구성된다. 프리젠테이션 계층은 사용자의 질의 검색어와 색인 구축기를 통해 구성된 색인 데이터베이스의 주제어를 비교하여 같으면 해당 정보를 HTML 문서 형태로 표현하여 사용자에게 보여준다. 그리고 트리거 계층은 로봇 에이전트와 연동하여 상호구동적(interactive)으로 동작하는 계층이다. 트리거 계층은 사용자의 질의 검색어가 색인 데이터베이스의 주제어로 존재하지 않을 경우 로봇 에이전트에게 URL 데이터베이스의 등록 도메인 이름을 이용하여 구한 검색어와 카테고리를 매개변수로 웹에서 다시 관련된 내용을 검색하여 색인 데이터베이스를 재구성하도록 요구한다.

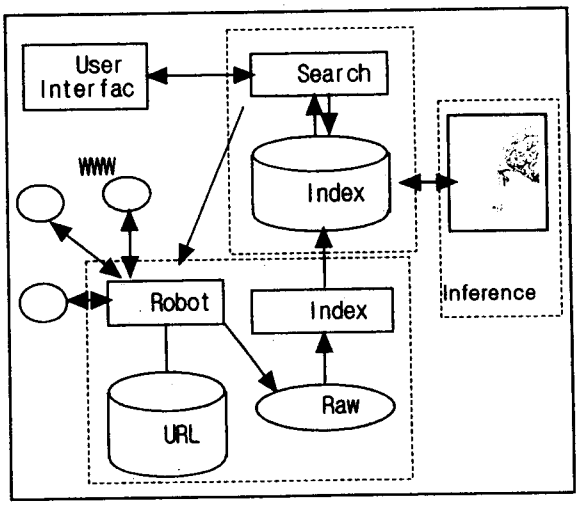

그림 4. 검색엔진의 내부 질의어 처리 구조

Fig. 4 Processing structure of internal query of search engine

3. 연관규칙 기반의 추론엔진 모듈의 구현

연관 규칙 탐사 기법을 구현하기 위해 제 1 단계에서는 사용자가 미리 정의한 최소 지지도를 만족하는 데이터 항목 집합을 탐사하는 단계로써 각 각의 데이터 항목에 대하여 지지도를 계산한 후 최소 지지도를 만족하는 데이터 항목들만 추출하다. 제2단계에서는 1단계에서 추출된 빈발 항목 집합들 중에서 사용자가 정의한 최소 신뢰도를 만족하는 규칙들을 탐사하여 최종 대상을 결정한다.

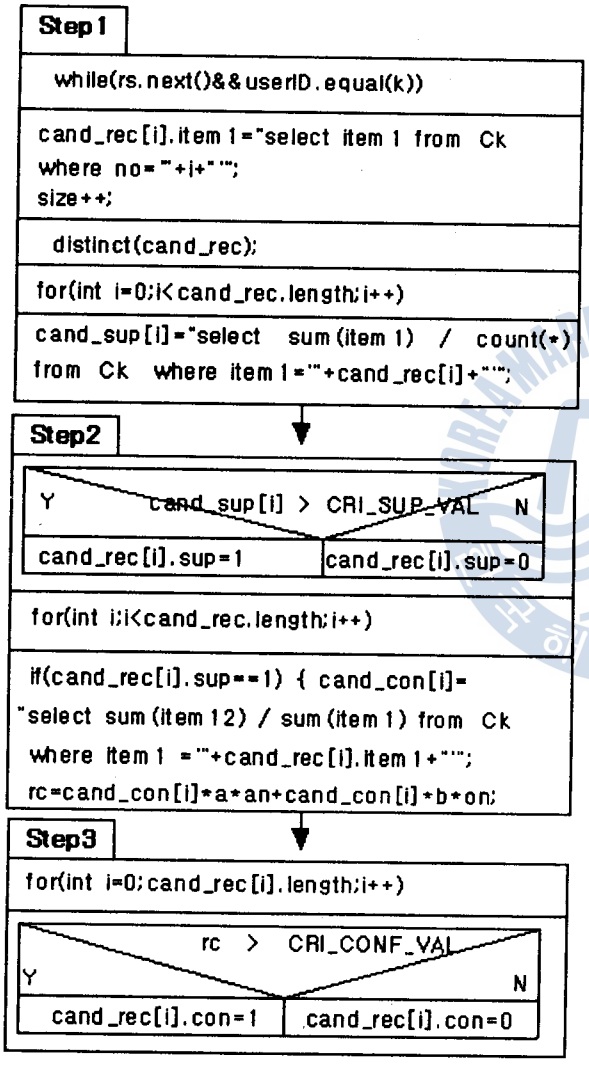

그림 5. 빈발항목집합 및 연관규칙 생성
Fig. 5 Generation of frequent item set and association rule

본 연구에서는 〈그림 5〉와 같이 빈발 항목 집합을 생성하기 위해 검색 창에서 주어진 질의어를 차 테이블의 기본키와 비교하여 일치하는 레코드들을 객체배열에 저장하고 지지도 평가함수를 이용하여 각 레코드 항목의 지지도를 계산하며 임계값으로 주어진 최소지지도를 만족하는 빈도 항목 집합을 결정한다. 이때 트랜잭션의 크기와 개수를 줄이기 위해 전체 데이터베이스를 탐색 대상으로 하지 않고 해당 사용자의 탐색 패턴이 저장된 로컬 데이터베이스로 탐색 영역을 제한한다. 또한 연관규칙을 효율적으로 생성하기 위해 1단계에서 추출된 빈도항목 집합을 대상으로 신뢰도평가함수를 통해 각 빈도항목의 신뢰도를 계산하고 임계값으로 주어진 최소신뢰도에 따라 연관규칙을 생성하여 질의어와의 연관성을 의미론적으로 해석 가능한 지능적 정보검색엔진을 구현하였다.

IV. 실험 및 고찰

본 연구에서 구현한 "AI-SEA"라는 해양관련 정보 전문 검색엔진은 JSP 언어를 사용하여 웹 검색에이전트와 웹 로봇에이전트 및 추론엔진 부분을 구성하였다. 색인 데이터베이스는 오라클 8.1을 사용하여 JDBC를 통해 연동하였다. 〈그림 6〉은 AI-SEA의 질의 결과 화면으로 질의어를 텍스트 필드에 입력하면 질의어와 매치(match)된 키워드를 추론엔진에서 연관규칙으로 추론하여 색인 데이터베이스에서 관련 정보를 찾아 하이퍼텍스트 문서로 결과를 보여주고 있다.

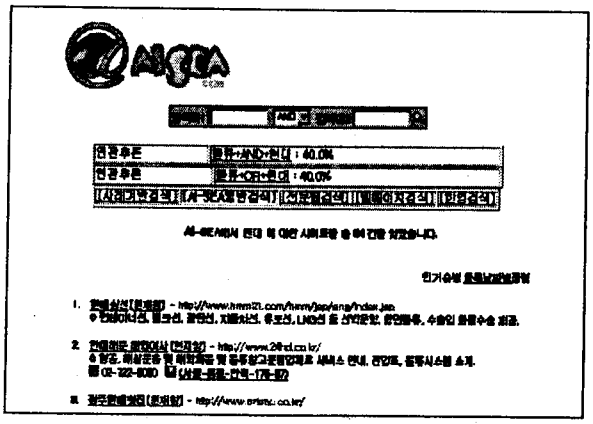

그림 6. AI-SEA 검색엔진의 질의 결과 화면
Fig. 6 Result of query of AI-SEA search engine

〈표 2〉는 질의어1과 질의어2간의 개별 연산에 대해서 S1항을 기준으로 발생한 S2항, S3항, S4항간의 신뢰도를 표시하고 최소신뢰도로 주어진 임계값 30%를 만족하는 경우의 연관규칙만을 채택하여 사용자에게 연관규칙 정보로 제공된다. 〈그림 7〉은 각 질의어들간의 관계 연산의 발생에 따른 신뢰도 변이값을 도식화한 것이다. 이때 신뢰도의 변화 추이는 질의어 상호간의 연산 결과에 따라 관련성 있는 질의어들간의 확신도는 증가하고 있음을 보여주고 있다. 따라서 해당 분야에 대한 전문지식이 부족한 일반 사용자가 하나의 질의어만을 입력해도 그와 관련성이 높은 질의어를 동반함으로써 보다 부가가치 높은 정보를 제공받을 수 있다.

표 2. 항목1(S1)을 기준으로 한 신뢰도 평가 예

Table. 2 Example of confidence evaluation about a item1(S1)

항목2 / 연산	S2		S3		S4	
	신뢰도	채택	신뢰도	채택	신뢰도	채택
S1∧S2, S1∧S3, S1∨S4	33.3%	√	33.3%	√	16.6%	×
S1∨S2	37.5%	√	25%	×	12.5%	×
S1∧S3	30%	√	40%	√	10%	×
S1∨S3	25%	×	41.7%	√	8.3%	×

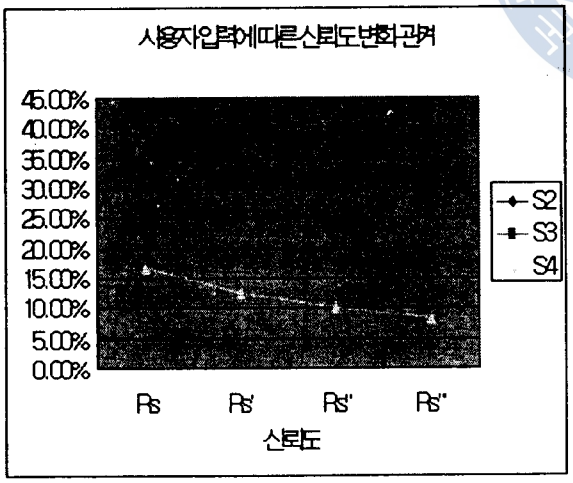

그림 7. S1항을 기준으로 한 신뢰도 평가 그래프

Fig. 7 Graph of confidence evaluation about a S1 item

〈표 3〉은 연관규칙탐사 방법, 전문검색엔진 방법, 상용검색엔진 방법에 따라 해양관련 주제 30개의 질의를 사용하여 검색을 수행하고 나온 결과를 관련 그룹별로 6개씩 묶어 질의 결과에 대한 평균치를 정확률과 재현률로 나타내었다. 상용검색엔진의 재현률을 나타내지 않은 이유는 검색에 이용된 범용검색엔진에서 보유하는 데이터베이스의 관련 문서 개수를 산출할 수 없기 때문이다.

표 3. 검색방법에 따른 정확률과 재현률

Table. 3 The correctness rate and the recall rate followed the search method

Group	연관규칙 탐사		전문검색 엔진		상용검색 엔진
	정확률	재현률	정확률	재현률	정확률
1	0.89	0.94	0.92	0.88	0.73
2	0.87	0.93	0.91	0.86	0.75
3	0.88	0.91	0.93	0.90	0.76
4	0.89	0.92	0.90	0.86	0.77
5	0.86	0.93	0.91	0.87	0.70

실험결과 연관규칙탐사기법이 적용된 검색이 평균적으로 전문검색엔진에 비해 재현률이 높고 상용검색엔진에 비해서는 정확률이 높은 것으로 나타났다.

V. 결론

검색엔진에 대한 유용성은 주제어에 대한 관련자료의 재현률과 정확률에 따라 평가된다. 일반적으로 상용검색엔진은 사용자 질의어에 대한 재현률은 높은 편이나 정확률은 낮은 경향을 나타내고 있고 전문검색엔진은 한정된 데이터베이스 정보를 제공함으로써 정확률은 상대적으로 높으나 재현률은 낮아서 사용자의 질의 요구를 충분히 만족시키지 못하고 있다. 따라서 본 연구에서는 검색엔진에서 재현율과 정확률을 동시에 높이기 위하여 사용자가 입력한 질의어와 연관된 정보를 함께 제공 할 수 있도록 연관규칙 탐사 기법을 적용하여 정확률 뿐만 아니라 재현률의 정도를 높일 수 있는 방법을 제안하여 전문검색엔진에 비해서는 재현률이 높고 상용검색엔진에 비해서는 정확률이 높은 결과를 확인하였다.

그러나 본 연구는 일반 사용자가 제공한 질의어를 기반으로 연관규칙을 추출함으로써 부정확한 연관규칙을 생성할 가능성이 있다. 따라서 이러한 문제를 검증하기 위해 전문가시스템에 대한 연구가 계속 필요할 것으로 사료된다.

참고문헌

[1] 고경자, 김인철, "사용자 접근 패턴 분석을 이용한 적응형 웹사이트 구축에 관한 연구", 한국지능정보시스템학회 추계학술대회, 2000.

[2] 서성보외3, "전자상거래에 적용가능한 시간연관규칙탐사기법", 한국정보과학회 학술논문지 vol. 26. no2, 1999.

[3] 장재정 · 오경환, "사용자의 피드백을 통한 퍼지 연관규칙의 웹사용자 마이닝", 한국정보과학회 학술논문지 vol.28.no2, 2001.

[4] 김정자 · 이도헌, "서열분석을 위한 연관규칙 탐사", 한국정보과학회 학술논문지 vol.28.no1, 2001.

[5] 김민정 · 박승수, "웹사이트 구조개선을 위한 웹페이지 연관규칙발견과 웹사이트 성능평가", 한국정보과학회 학술논문지 vol.28.no2, 2001

[6] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, "Database mining : A performance perspective", IEEE Transactions on Knowledge and Data Engineering, December 1993.

[7] 이정원외 6, "데이터마이닝 알고리즘의 분류 및 분석", 정보과학회논문지 28호 3권, 2001.9

[8] 권경희 · 정균락, "연관규칙탐사를 위한 효율적인 자료구조", 한국정보과학회 학술논문지 vol. 28. no2, 2001

저자소개

하 창 승

1984. 2 한국해양대학교 항해학과 졸업(공학사)

1992. 2 한국해양대학교 전자통신공학과(공학석사)

2001. 2 한국해양대학교 전자통신공학과(공학박사수료)

1996. 9 ~ 현재 동명대학 정보통신계열 조교수

윤 병 수

1998. 2 한국해양대학교 제어계측공학과 졸업(공학사)

2001. 8 한국해양대학교 컴퓨터공학과(공학석사)

2002. 3 한국해양대학교 컴퓨터공학과 박사과정 재학 중

류 길 수

1976. 2 한국해양대학교 기관학과 졸업(공학사)

1979. 2 한국해양대학교 대학원 기관학과(공학석사)

1986 일본동경공업대학 대학원(공학석사-정보공학)

1989 일본동경공업대학 대학원(공학박사~정보공학)

1982 ~ 현재 한국해양대학교 기계정보공학부 교수