

工學博士 學位論文

연관규칙 탐사와 사례기반 추론 기법에 의한  
유무선 통합 지능형 검색 에이전트  
시스템의 구현에 관한 연구

A Study on the Implementation of Convergence Intelligent Search Agent System  
using Association Rules and Case-Based Reasoning Techniques

指導教授 柳 吉 洙

2004年 2月

韓國海洋大學校 大學院

電子通信工學科 河 昌 昇

# 목 차

<b>제 1 장 서론</b> .....	1
1.1 연구의 배경 .....	1
1.2 연구의 목적 및 내용 .....	3
<b>제 2 장 범용검색엔진 및 지능형 검색 에이전트의 고찰</b> .....	6
2.1 범용검색엔진의 구현 기술 .....	6
2.1.1 검색 로봇 .....	7
2.1.2 인덱서 .....	10
2.1.3 질의서버 .....	14
2.2 지능형 에이전트의 고찰 .....	15
2.2.1 검색 에이전트의 지능적 학습 방법 .....	16
2.2.2 지능형 에이전트들의 특성 .....	18
<b>제 3 장 지능형 웹 검색을 위한 이론적 연구</b> .....	27
3.1 KDD 관련 기술 .....	27
3.2 연관규칙 탐사 기법 .....	29
3.2.1 빈발항목집합의 정의 .....	31
3.2.2 연관규칙의 정의 .....	32
3.2.3 연관규칙 탐사의 접근 방식 .....	34
3.3 사례기반 추론 .....	40
3.3.1 사례의 구성과 모형의 표현 .....	41
3.3.2 사례기반 추론 과정 .....	44

3.3.3 유사도 평가 .....	46
3.4 지능형 검색을 위한 타당성 검토 .....	48
<b>제 4 장 질의어 처리 알고리즘의 설계 .....</b>	<b>51</b>
4.1 연관규칙 탐사 알고리즘 설계 .....	51
4.1.1 빈발항목 생성 .....	51
4.1.2 신뢰도 평가함수 .....	53
4.2 사례기반 추론 알고리즘 설계 .....	55
4.2.1 사례기반 추론 절차 .....	55
4.2.2 사용자 모델링을 통한 사례기반 학습 .....	57
4.2.3 유사도 평가 방법 .....	58
<b>제 5 장 지능형 에이전트 시스템의 구현 .....</b>	<b>62</b>
5.1 지능형 에이전트 시스템의 구성 .....	62
5.2 로봇부 및 검색부 .....	64
5.2.1 로봇부의 구조 .....	64
5.2.2 검색부의 구조 .....	68
5.3 연관규칙 추론부의 구조 .....	73
5.4 사례기반 추론부의 구조 .....	77
5.5 무선 인터넷 연결 모듈 .....	82
5.5.1 유무선 통신의 접근 위상 .....	82
5.5.2 무선 인터넷 구현 기술 .....	84
5.5.3 무선 에이전트의 구현 및 처리 절차 .....	87
5.5.4 무선 인터넷 접속 및 검토 .....	90

<b>제 6 장 실험 및 고찰</b> .....	92
6.1 연관규칙 탐사 기법을 통한 실험 .....	92
6.1.1 실험성능 평가 방법 .....	92
6.1.2 실험 환경 및 결과 분석 .....	93
6.2 사례기반 추론 기법을 통한 실험 .....	95
<b>제 7 장 결 론</b> .....	101
<b>참고문헌</b> .....	103
<b>부 록</b> .....	107

## 표 목 차

표 2.1 단어와 문서의 배열(비트맵 색인) .....	11
표 2.2 축 변환 후 배열(비트벡터) .....	12
표 2.3 각 시스템별 관련도 평가함수 .....	25
표 3.1 사례기반 추론의 하부 항목 .....	42
표 3.2 지능형 알고리즘 구현을 위한 타당성 검토 항목 .....	49
표 4.1 그림 4.3을 이용한 신뢰도 평가 예 .....	55
표 5.1 트랜잭션 테이블 .....	73
표 5.2 발견된 연관규칙 .....	74
표 5.3 사례베이스의 스킴 구조 .....	81
표 5.4 무선 인터넷의 단계별 처리 기능 .....	86
표 6.1 연관규칙 탐사 기법의 검색방법에 따른 정확율과 재현률 .....	93
표 6.2 질의어1(항S1)을 기준으로 한 신뢰도 평가 .....	94
표 6.3 사례기반 추론 기법의 검색방법에 따른 정확율과 재현률 .....	97

## 그림 목 차

그림 2.1 범용검색엔진의 구성요소 .....	6
그림 2.2 검색 로봇의 탐색공간 그래프 .....	8
그림 2.3 검색 로봇의 동작 과정 .....	10
그림 2.4 질의 처리 모듈 .....	15
그림 2.5 Personal WebWatcher의 정보검색 모델 .....	19
그림 2.6 Letizia의 정보검색 모델 .....	21
그림 2.7 WiseWire의 정보검색 모델 .....	23
그림 3.1 연관규칙을 위한 지지도 및 신뢰도 .....	30
그림 3.2 빈발항목 선정과 지지도계산 .....	36
그림 3.3 사례의 프레임 표현 과정 .....	42
그림 3.4 사례기반 추론 모형의 개념적 표현 .....	43
그림 3.5 사례기반 추론 흐름도 .....	44
그림 3.6 두 패턴 벡터간 코사인 값을 통한 유사도 평가 .....	47
그림 4.1 그룹화 규칙 생성 알고리즘 .....	52
그림 4.2 그룹화를 통한 항목조합 지지도 계산 .....	53
그림 4.3 카티전 프로덕트를 통한 신뢰도 계산 .....	54
그림 4.4 인지적 확률모델 기반의 사례추론 알고리즘 .....	56
그림 4.5 사용자 모델링을 통한 사례기반 학습 과정 .....	58
그림 5.1 AI-SEA 시스템의 질의 결과 화면 .....	63
그림 5.2 AI-SEA 시스템의 개념적 구조 .....	63
그림 5.3 로봇부의 개념적 구조 .....	67
그림 5.4 로봇부의 세부 동작 과정 .....	67
그림 5.5 로봇부의 색인데이터베이스 구현 .....	68

그림 5.6	검색부의 색인데이터베이스 연동 모델 .....	69
그림 5.7	검색부의 색인데이터베이스 연동 코드 .....	71
그림 5.8	색인 데이터베이스 및 관련 데이터베이스의 릴레이션 구조 .....	72
그림 5.9	연관규칙 추론부의 구조 .....	75
그림 5.10	최소 지지도 및 최소 신뢰도 임계값 지정 .....	75
그림 5.11	빈발항목집합 및 연관규칙 생성 코드 .....	76
그림 5.12	연관규칙 탐사를 통해 연관어가 추출된 검색 결과 .....	77
그림 5.13	사례기반 추론부의 구조 .....	78
그림 5.14	사례별 카테고리 그룹 계층도 .....	79
그림 5.15	유사도 평가 모듈의 구현 코드 .....	81
그림 5.16	사례기반 추론 결과 화면 .....	82
그림 5.17	무선통신을 통한 유선 인터넷 접근 위상 .....	84
그림 5.18	WAP방식의 무선 클라이언트/서버 통신 모델 .....	85
그림 5.19	WAP 프로토콜 스택 구조 .....	86
그림 5.20	무선 인터넷 망을 통한 검색에이전트 연동 모델 .....	88
그림 5.21	무선 인터넷 정보 흐름도 .....	89
그림 5.22	무선단말기(018)에서의 검색 과정 .....	90
그림 5.23	무선단말기(019)에서의 검색 과정 .....	91
그림 6.1	연관규칙 결정을 위한 두 개의 질의어의 불린 연산 .....	94
그림 6.2	표 6.2에 대한 신뢰도 변이 그래프 .....	95
그림 6.3	질의어 입력 초과 검색 종류 .....	96
그림 6.4	사례기반검색에서 정확율과 재현률의 변화 추이 .....	98
그림 6.5	전문검색엔진을 통한 정확율과 재현률의 변화 추이 .....	98
그림 6.6	사례기반검색과 전문검색의 정확율과 재현률의 변화 추이 .....	100
그림 6.7	전문웹검색과 웹페이지 검색에서의 정확율 변화 추이 .....	100

## **Abstract**

Recently, since variety of information through the web is excessively expanding, people are not only depending on searching engines, but also, spending more time for searching they need. In order to improve efficiency of the searching engine, recently, intelligent agents based on preference of each user are being developed. However, some intelligent agent depending on the very limited assumptions for the considering users preference produces almost irrelevant results.

Therefore, the intelligent agent is necessary to support a personalized searching function for the personal and active evaluation of the users preference. It also should support semantically related information by analyzing correlation of the information.

In this thesis, association rules and case-based reasoning method such that can analyze actions and characteristics of each user and extract individual behavior pattern rules by repetitive learning for supporting reasoning are suggested. The association rules and case-based reasoning method used in the data mining field can measure the degree of verification of the related data depending on the supportability and reliability, and can evaluate the degree of the relevance by its evaluation. Therefore, the recall ratio and the precision ratio can be improved by applying these rules for verifying the relevance among web documents.

The grouping rule formation algorithm for the relevance rule searching method and a recognized probability model for the implementation of the case-based reasoning method is presented. The



former calculates supportability and reliability to measure the verification ratio, and the latter evaluates relevance by measuring the relativity.

The learning through the user modeling properly feedbacks the cleaned information of case presented level, and have expanded and adaptive function of the knowledge by persistently changing the users category group.

Moreover, in the study on the accessibility of the wired internet information, the advantages of improving the accessibility and the mobility can be verified by applying WAP technology on the web-searching agent. As a result of the performance evaluation, the association rules method showed better recall ratio than the expert searching engine and showed higher precision ratio than the general one. The case-based searching method showed better precision than both in the expert and the general one.

Therefore, the suggested searching engine using the association rules and the case-based reasoning method statistically based on the users searching behavior showed superior results both on the recall ratio and the precision ratio compare to the any other existing searching engines.

# 제 1 장 서 론

## 1.1 연구의 배경

최근 기업, 단체 및 개인에 이르기까지 많은 정보 제공자들이 웹을 통하여 정보를 제공함에 따라 웹을 통한 정보의 검색 또한 점점 일반화되고 있고 웹 상의 정보 양도 급속히 증가하고 있다. 현재 웹 상에서 색인 가능한 문서의 수가 25억 개를 넘었고, 동적으로 생성되는 웹 페이지의 수는 5,500억개 정도이며, 하루에 새롭게 생겨나는 웹 문서도 백 만개를 넘어서고 있다.<sup>[1]</sup> 정보검색에서 정보량의 급속한 증대는 자신에게 유용한 정보를 찾는데 점점 더 많은 시간을 투자해야 하고 범용검색엔진의 문서 처리 시간을 지연시키고 있다. 문서의 양적 증가로 인한 처리지연 외에도 웹 문서의 분류나 표현규칙이 아직 표준화되어 있지 않아 특정 주제에 대한 사용자의 정보 요구를 정확하게 인식하지 못해 사용자의 선호도나 검색 목적에 따라 개인화된 정보를 제공하지 못하는 질적인 문제도 있다.<sup>[2]</sup>

범용검색엔진은 수많은 웹 문서로부터 다양한 주제와 풍부한 관련 정보를 제공할 수 있는 이점이 있지만 자원발견 단계, 정보추출 단계, 색인화 단계, 정보유지 단계, 정보제공 단계 등에서 다음과 같은 문제점이 있다.<sup>[3][4]</sup> 먼저 자원발견 단계에서는 급증하는 웹 문서를 수용하기 위해 발견적(heuristic)인 탐색방법을 사용하지 않고 너비우선탐색법과 같은 단순한 방법으로 링크 그래프를 방문하기 때문에 검색 효율이 낮고 반복적으로 동작하는 검색 로봇은 네트워크 트래픽을 증가시킨다. 정보추출 단계에서는 의미망(semantic network)을 구성하지 못한 색인데이터베이스는 사용자에게 유용한 핵심 지식의 창출 및 개인별 정보 가공을 하지 못한다. 또한 색인화 단계에는 단어의 출현빈도(TF : term frequency)에만 근거하여 단순빈도와 상대빈도를 구분하지 않고 역문

헌빈도(IDF : inverse document frequency)도 고려하지 않으므로 색인어 분류의 정확도를 향상시키지 못하고 있다. 정보유지 단계에서는 동적으로 변하는 웹의 성질 때문에 웹 문서들을 주기적으로 검색해서 변화된 정보를 색인데이터베이스(index database)에 반영하는 검색 주기의 즉시성도 낮은 편이다. 정보제공 단계에서는 주어진 질의어와 색인데이터베이스의 색인어들을 단순 패턴 비교를 통해 일치하는 정보를 검색하는 기법을 사용함으로써 검색 효율이 매우 낮다.

이러한 문제를 부분적으로 해결하기 위해 특정 영역에 관한 세부적이고 전문적인 내용이나 전문용어 혹은 이와 유사한 용어에 대해 특화된 정보를 제공하는 전문검색엔진이 개발되고 있다. 전문검색엔진은 특정분야의 전문지식을 디렉토리 서비스 형태나 로컬 데이터베이스 기반의 웹 서비스형태로 제공한다. 분야별 전문검색엔진의 다양화는 외국에서는 이미 대중화된 인터넷 서비스 중 하나로 미국만 해도 1,800여 개의 전문검색엔진이 사용되고 있다.<sup>[5][6]</sup> 예를 들어 뉴스의 헤드라인만 검색해 주는 사이트(www.moreover.com), 연방법과 정부의 웹사이트만 전문적으로 검색해 주는 사이트(www.findlaw.com), 과학기술과 관련된 정보만 제공하는 사이트(www.biolinks.com) 등으로 특성화되어 가고 있고 국내에서도 교육정보만을 전문적으로 제공해주는 에듀인포 사이트(eduinfo4k.wo.to)가 있다. 그러나 전문검색엔진에서 제공되는 정보는 질의어와 관련성이 깊고 주제별로 집약된 정보이지만 정보의 양과 수준은 전문지식을 필요로 하는 사용자의 기대보다 대부분 미흡한 경우가 많다. 특히 에듀인포 사이트는 신뢰성 검사기능과 사용자 프로파일에 기초한 정보 필터링 기능이 부족하여 상황에 따른 선별적 추출기능이 미약한 편이다.<sup>[7]</sup> 이것은 검색엔진이 해당 질의어에 대해 디렉토리 검색이나 로컬 데이터베이스만을 검색함으로써 한정된 영역의 정보를 사용하고 질의어와 연관성 있는 정보도 함께 제공하지 못하는 정보획득 및 추출의 한계성 때문이다.<sup>[8]</sup>

## 1.2 연구의 목적 및 내용

거대한 가상공간을 대상으로 하는 정보검색에서는 신속한 검색이나 풍부한 자료의 제공 못지 않게 사용자의 의도를 정확히 파악하여 개인화된 전문 지식을 제공하거나 사용자의 선호도를 고려하는 지능형 에이전트가 필요하다. 지능형 에이전트는 문서의 의미를 분석하여 카테고리 그룹을 분류하거나 특정한 응용영역에 대한 경험적 지식을 재활용하는 학습(learning) 기능을 제공하고 검색을 요청하는 사용자의 선호도에 따라 다른 검색 결과를 제시한다. 이러한 에이전트의 지능적 행위는 범용검색엔진과 전문검색엔진의 문제점과 한계성을 극복하는 대안이 될 수 있다.

사용자의 관심도를 정보 검색에 이용하는 지능형 에이전트 기술은 여러 분야에서 활발히 연구되고 있지만 웹 에이전트 연구의 주요 모델이 되는 시스템은 다음과 같다. 먼저 Personal WebWatcher 시스템은 TF-IDF와 베이지안 확률을 이용한 학습방법을 이용하여 사용자가 관심을 가지는 문서들의 형태와 내용에 대한 개념 모델을 만들고 사용자가 정보 검색 시에 이 개념 모델을 이용하여 관심 문서를 예측하여 제안한다.<sup>[9][10]</sup> Letizia 시스템은 클라이언트형 에이전트로 기계학습 접근법을 이용하여 사용자의 행동을 관찰하고 수집된 행동 패턴을 분석하여 무감독 학습방법을 통해 사용자의 관심 문서를 예측한다.<sup>[11]</sup> WiseWire는 감독 학습방법과 무감독 학습방법을 함께 사용하는 혼합형 학습방법을 지니며 보조적으로 다른 사람의 행동 패턴도 함께 고려하는 협력학습(collaborative learning) 기능도 제공한다. 또한 행동 패턴을 묵시적으로 학습하는 방법외에 사용자에게 문서의 관심 정도를 0에서 10까지 세분화된 수치값을 입력하도록 하는 명시적 학습방법도 제공하고 있다.<sup>[12]</sup> SMART 시스템은 벡터공간모델 이론에 기초하여 군집파일(clustered file)을 색인으로 사용하여 사용자의 질의 유사도를 측정하고 있으며<sup>[13]</sup> GLOSS 시스템은 정보저장소의 위치를 지시해 주는 메타데이터를 이용하며 문서빈도나 용어가중치의 합과 같은

통계치를 기반으로 관련도를 측정한다.<sup>[14]</sup> SavvySearch 시스템은 강화학습방법(reinforcement learning method)으로 관련성을 측정하고 분산된 환경에서 능동적 적응성을 특성으로 가지고 있다.<sup>[15]</sup> Amalthea 시스템은 복잡한 영역의 문제 해결을 위하여 여러 에이전트가 상호 협동하는 멀티에이전트 시스템으로 구성되며 관련성 측정을 위해서 진화론(ecosystem)적 학습방법이 이용되고 있다.<sup>[16]</sup>

그러나 SMART 시스템과 GIOSS 시스템 같은 일부 지능형 에이전트들은 사용자별 선호도를 고려할 때 매우 제한적인 가정(assumption)에 의존하여 검색함으로써 관련성 낮은 결과를 제공하거나 사용자별로 개인화된 정보를 제공하지 못하는 경우가 종종 있다. 따라서 본 논문에서는 사용자별로 개인화되고 연관성 있는 정보를 제공하기 위해 웹 검색 에이전트에 데이터마이닝의 연관규칙 탐사(association rules) 기법과 사례기반 추론(case-based reasoning) 기법의 적용을 제안한다.

연관규칙탐사 기법은 항목집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 찾아 데이터간의 관계 규칙을 발견하는 기법이며 사례기반 추론은 주어진 문제를 해결하기 위해 과거의 유사한 사례를 문제의 상황에 맞게 응용하여 새로운 해를 발견하는 기법이다. 연관규칙 탐사 알고리즘의 구현을 위해 Apriori 알고리즘을 검색엔진의 특성에 맞게 변형한 ‘그룹화 규칙 생성(grouping rule formation) 알고리즘’을 개발하였다. 이 알고리즘은 순서적 의미를 갖는 두 요소 항목 집합의 조합에서 첫 요소 항목을 빈발항목 생성과 지지도 계산에 기준 요소로 선정하여 첫 요소 항목을 중심으로 그룹화하여 연관규칙의 신뢰도를 평가한다. 사례기반 추론 알고리즘의 구현은 ‘인지적 확률모델(recognized probability model)기반의 사례추론 알고리즘’을 개발하여 주어진 질의어의 관련 정보 카테고리를 결정하도록 하였다. 인지적 확률모델이란 질의어를 의미론적으로 해석하기 위해 확률적으로 관심 빈도가 높은 카테고리

에 속하는 색인어가 그 질의어와 관련성이 높다는 이론에 기초한다.

연관규칙 탐사와 사례기반 추론 기법은 지지도와 신뢰도에 따라 연관자료의 확신도를 측정하고 유사도에 따라 관련성 정도를 평가하기 때문에 기존의 검색 방법에 비해 자료의 재현률과 정확율을 개선할 수 있다. 지지도와 신뢰도의 평가는 불린 연산에 의해 축적된 연관지식을 이용하고 유사도 측정은 사용자의 과거 검색행위와 관련된 경험적 지식을 사례로 활용한다. 연관규칙 탐사와 사례기반 추론 기법이 적용된 에이전트에서는 학습과 추론 기능을 통해 관련성 높은 카테고리 그룹의 정보를 제공하며 질의어와 관련된 연관어도 함께 제공한다.

다른 전문검색엔진이나 범용검색엔진과의 성능평가 실험에서는 본 논문에서 개발한 AI-SEA 시스템이 적용 기법에 따라 차이는 있지만 전문검색엔진에 비해 정확율과 재현률이 높고 범용검색엔진에 비해서는 정확율이 높은 결과를 나타내었다.

본 논문은 7장으로 구성된다. 2장은 일반적인 범용검색엔진의 구성 및 지능형 검색 에이전트에 대해 살펴보고 3장은 지능적 웹 검색을 위한 이론과 방법에 대해 기술한다. 4장은 본 논문에서 구현한 시스템의 지능적 추론을 위한 질의어 처리 알고리즘을 설계하고 5장은 지능형 에이전트 시스템의 구현 방법에 대해 기술한다. 6장은 실험 결과에 대해 고찰하고 7장은 결론으로 끝을 맺는다.

## 제 2 장 범용검색엔진 및 지능형 검색 에이전트의 고찰

정보검색 에이전트는 사용자를 대신하여 인터넷에 산재해 있는 정보들에 대하여 정보검색과 정보여과 등의 기능을 자동적으로 수행하는 소프트웨어를 말한다. 정보검색 에이전트가 지니는 공통적인 특징은 자율성과 협동성, 그리고 적응성이다. 자율성은 사용자의 직접적인 개입 없이 자율적으로 행동하는 성질을 말하며 협동성은 주어진 작업을 수행하기 위해 다른 에이전트나 사용자와 협력하는 성질을 의미한다. 적응성은 변화하는 환경에 따라 에이전트가 학습할 수 있는 능력을 의미하며 에이전트의 수행 효율 정도를 평가하는 기준이 된다. 따라서 자율성이나 협동성, 적응성 등의 성질을 만족하지 못하는 범용검색엔진은 에이전트라고 볼 수 없다.

### 2.1 범용검색엔진의 구현 기술

인터넷상에 산재하는 정보를 검색하는 범용검색엔진은 일반적으로 그림 2.1과 같이 검색 로봇, 인덱서(indexer), 질의서버(query server)로 구성된다.<sup>[7]</sup>

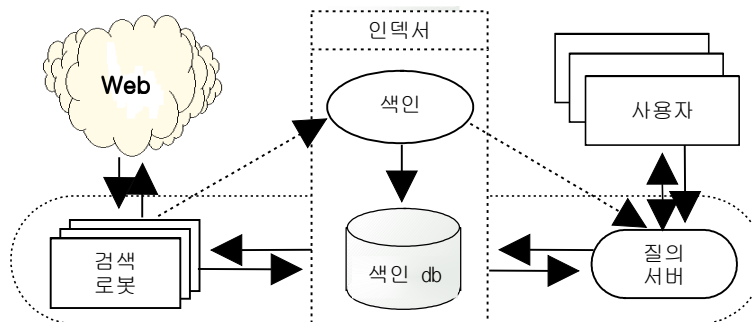


그림 2.1 범용검색엔진의 구성요소

### 2.1.1 검색 로봇

검색 로봇은 자동으로 웹 문서의 하이퍼링크 구조를 따라 다니며 문서를 추출하고 다시 그 웹 문서에서 참조되는 다른 웹 문서들을 순환적으로 탐색하여 관련 정보를 추출하는 프로그램이다. 검색 로봇은 인터넷에 있는 각각의 웹 서버에 직접 접근하여 웹 문서에서 URL을 추출하고 페이지들을 모아 인덱스에 전달한다.

#### 1) 탐색공간 항해전략

검색 로봇의 기본 동작은 정보를 추출하기 위해 초기 URL로 주어진 특정한 웹사이트로부터 시작하여 하이퍼링크를 따라 연결된 다른 모든 웹 페이지들을 방문한다. 하나의 페이지가 선정되면 페이지내의 모든 링크주소들을 큐(queue) 속에 저장하고 다음 URL을 큐에서 선택하여 웹 페이지를 가져오는 작업을 반복한다. 이때 하나의 문서를 기점으로 그 문서에 연결된 여러 링크를 어떤 순서로 방문할지 고려해야한다. 이것을 항해전략(traversal strategies)이라 하며 일반적으로 항해전략은 그림 2.2와 같은 사이트 맵에서 (V1, V2, V4, V8, V5, V6, V3, V7)순으로 링크를 탐색하는 깊이우선 탐색과 (V1, V2, V3, V4, V5, V6, V7, V8)순으로 링크를 탐색하는 너비우선 탐색 방법이 있다.

깊이우선 탐색은 특별한 전략이 없다면 하나의 사이트에 너무 오래 머물게 되어 넓은 영역의 정보 수집이 어려울 뿐만 아니라 한 서버의 URL만을 계속적으로 접근하여 그 서버에 부하를 가중시키거나 기능을 마비시키기도 한다. 따라서 대부분의 검색 로봇이 너비우선 탐색 방법을 선택한다. 이외에도 발견적 탐색 방법 중의 하나인 최적우선 탐색 방법이 고려될 수 있다. 이 방법은 경험적 탐색기법을 통해 URL의 길이를 비교하여 길이가 짧은 링크를 우선적으로 방문한다. 이것은 URL의 길이가 짧을수록 특정 사이트의 상위 레벨의 위치를 나타낼 가능성이 높으므로 좀더 넓은 영역의 정보수집이 가능하기 때



문이다.

항해전략에 따라 약간의 차이점은 있지만 범용검색엔진의 공통적인 문제점은 단순질의에 대한 검색결과가 지나치게 많고 정확도가 떨어진다는 것이다. 또한 검색 로봇의 근본적인 로봇문제(robot problem)로서 로봇이 링크 URL의 특성을 인식하지 못하는 문제가 있다. 즉, 내부 데이터베이스 연동 URL이나 일시적인 링크에 대해서는 색인화 할 필요가 없지만 로봇은 이를 판단하지 못한다. 이를 해결하기 위해 'robots.txt'라는 파일을 이용한 로봇 배제 표준안(robot exclusion standard)이 제안되었다.<sup>[17]</sup> 이 파일은 웹서버에 대한 접근 정책정보를 미리 제공하여 로봇이 색인화할 필요가 없는 문서를 명시함으로써 검색 로봇이 불필요한 URL의 접근으로 인해 발생하는 속도 저하 문제는 방지 할 수 있지만 이 표준안을 따르지 않는 정보 사이트에는 적용되지 못하는 문제점은 여전히 남게 된다.

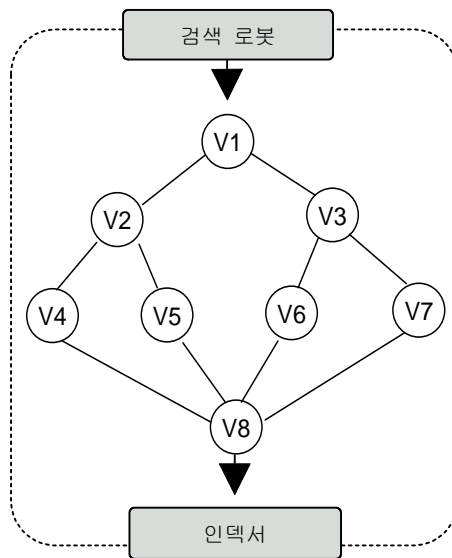


그림 2.2 검색 로봇의 탐색공간 그래프

## 2) 검색 로봇의 동작과정

검색 로봇은 그림 2.3과 같이 주기적으로 혹은 비 주기적으로 웹 공간을 자동 향해하면서 관련 정보를 수집한다. 검색 로봇의 동작 과정은 다음과 같다.

- 단계 1 : 초기에 방문할 URL 주소를 주고 검색 로봇을 실행한다.
- 단계 2 : 호스트 이름에 따라 'http://host\_name/robots.txt'에 접근한다.
- 단계 3 : 'robots.txt' 파일의 내용을 받는다.
- 단계 4 : 'robots.txt' 파일의 내용을 분석해 자신이 들어갈 수 있는 사이트인지 아닌지 확인한다.
- 단계 5 : 자신을 배제하는 사이트가 아니면 다시 사이트에 접근한다.
- 단계 6 : 접근한 사이트의 문서를 전달받아 임시 파일로 저장한다.
- 단계 7 : 저장된 임시 파일을 분석해 사이트의 URL을 추출하고 URL 테이블에 저장한다.
- 단계 8 : 저장된 임시 파일을 분석해 키워드를 추출하고 필요한 정보를 색인데이터베이스에 저장한다.
- 단계 9 : URL 테이블에서 다음에 방문할 URL을 받고 단계 2 또는 단계 5에서 단계 8까지의 과정을 반복한다.

URL 저장과 관련하여 단계 7에서 추출된 모든 URL은 상대 URL이 아닌 절대 URL로 변환해야한다. 즉, 웹 문서의 링크 페스인 '/dir/index.html'과 같은 상대 URL은 'http://host\_name/dir/index.html'과 같은 절대 URL로 바꾼다. 추출한 URL은 나중에 참조할 수 있도록 URL 테이블에 저장하며, 이때 몇 개의 URL을 추측해 함께 저장하기도 한다.

예를 들어 'http://host/dir/subdir/file.html'이 저장할 URL이라면 이것 외에도 'http://host/dir/subdir/', 'http://host/dir/', 'http://host/'도 내용을 갖고 있다

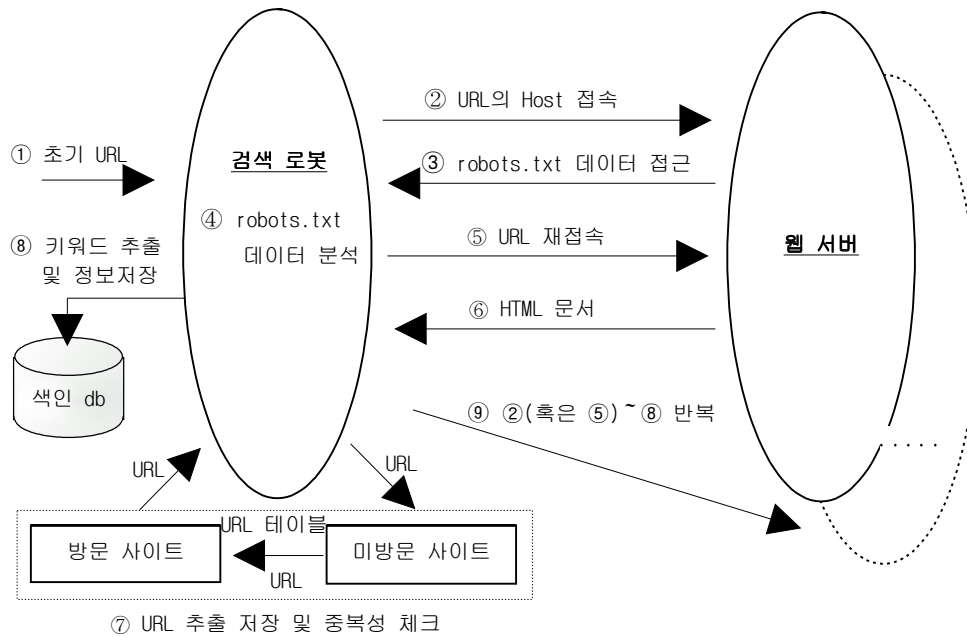


그림 2.3 검색로봇의 동작 과정

고 추측할 수 있다. 그리고 추출한 URL을 URL 테이블에 저장할 때는 이미 등록된 것인지 아닌지 검사해 중복되지 않도록 한다.

### 2.1.2 인덱서

인덱서는 검색 로봇이 추출한 정보를 색인데이터베이스에 저장하기 위해 한 단어(색인어)가 어떤 문서에 출현했는지를 신속하게 알 수 있도록 구조화하는 부분이다. 인덱서가 색인화하는 방법에는 비트맵 색인화, 역파일 색인화 (inverted file indexing), 요약파일 색인화(signature file indexing) 기법이 있지만 빠른 검색을 위해 역 파일 색인화 기법을 주로 사용한다.<sup>[18][19]</sup>

#### 1) 역파일 색인화

역파일 색인화란 비트맵 색인화 기법에서 좌표 변환된 역파일의 비트맵에서

값이 1인 항목만을 리스트 구조로 표현한 것이다. 비트맵 색인에서 축 변환된 배열을 저장한 파일을 넓은 의미에서 역파일이라고 부른다. 여기에서 사용되는 이진배열을 비트 벡터라고 한다. 비트 벡터는 가장 단순한 형태로 문서들을 표현하는 방법이며, 문서와 단어들을 배열 안의 위치정보만 가지고 찾아낼 수 있다는 점에서 가장 쉽게 프로그래밍 할 수 있는 구조이지만 배열의 크기가 너무 크고 추가 및 삭제연산을 수행할 때 비트벡터를 이동시켜야 하는 단점이 있다.

역파일 색인화 처리 과정은 다음과 같다.

- 단계 1 : 문서로부터 단어들을 추출하고 문서  $D_i$ 와 단어  $W_i$ 들을 다음 표 2.1과 같은 2차원 배열인 비트맵 색인으로 구성한다.

표 2.1 단어와 문서의 배열(비트맵 색인)

단어 \ 문서	$W_1$	$W_2$	$W_3$	...	$W_m$
$D_1$	0	1	0	...	1
$D_2$	0	0	1	...	0
$D_3$	0	1	1	...	0
...	...	...	...	...	...
$D_n$	1	0	0	...	0

- 단계 2 : 문서를 기준으로 추출된 단어는 색인 구조로 저장될 때 단어를 기준으로 저장하는 것이 활용도가 높으므로 표 2.2와 같이 배열의 가로와 세로축을 바꾸는 좌표축 변환 작업을 수행하여 비트 벡터를 생성한다.

표 2.2 축 변환 후 배열(비트벡터)

문서 단어	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	...	D <sub>n</sub>
W <sub>1</sub>	0	0	0	...	1
W <sub>2</sub>	1	0	1	...	0
W <sub>3</sub>	0	1	1	...	0
...	...	...	...	...	...
W <sub>m</sub>	1	0	0	...	0

- 단계 3 : 비트벡터에서 단어를 기준으로 단어가 출현된 문서의 상대적 위치값을 리스트형태로 표현한다.

$$W_1=(\dots, D_n)$$

$$W_2=(D_1, D_3, \dots)$$

$$W_3=(D_2, D_3, \dots)$$

$$W_m=(D_1, \dots)$$

- 단계 4 : 검색에 거의 영향을 미치지 않는 일반적인 단어들이 불용어(stopword)를 제거한다. 불용어의 제거는 색인의 크기를 큰 폭으로 줄일 수 있기 때문에 역파일 기법에서 불용어는 반드시 고려되어야 한다.
- 단계 5 : 각 단어가 각 문서에서 출현한 빈도 수를 산출하여 리스트 구조로 표현한다. 여기에 단어들이 문서 내에서 나타난 위치정보까지 표현하면 보통 상용 검색엔진에서 사용되는 정도의 정보표현 수준에 이를 수 있다.

$$W_1=(\dots, \{<D_n,3>128\})$$

$$W_2=(\{<D_1,8>16,169\}, \{<D_3,4>15,32,37,88,113,234,399\}, \dots)$$

$$W_3=(\{<D_2,6>33,185\}, \{<D_3,2>19,34,48,56,79\}, \dots)$$

$$W_m=(\{<D_1,4>11,80,109,146\}, \dots)$$

## 2) 색인화 알고리즘

대부분의 검색엔진은 로봇이 추출한 문서전체의 내용을 색인화하는 전문 색인화(full-text indexing) 방법을 선택하고 색인화는 특정 단어와 각 문서들의 관계로 연관성의 정도가 결정된다. 하나 이상의 문서가 같은 단어와 관계가 있을 경우 어느 문서가 더 관련성이 높은지 구별할 필요가 있다. 각 문서에 대한 특정 단어의 관련성 정도를 판별하는데는 주로 TF-IDF 알고리즘을 이용한다. 문서내 출현빈도(TF)는 한 단어가 한 문서 내에서 등장하는 횟수를 나타내고 문헌빈도(DF : document frequency)는 한 단어가 N개의 문서의 집합 중에서 몇 개의 문서에서 나타나는 가를 의미한다. 한 단어가 한 문서에서 여러 번 나타난다면 그 문서는 관련성이 높은 문서라고 판단 할 수 있지만 여러 문서에 걸쳐 나타난다면 그 단어에 대한 중요도는 낮다고 보아야 함으로 문헌빈도의 역인 역문헌빈도(IDF)가 이용된다.

따라서 문서의 우선순위는 문서 내 출현빈도 값과 역 문헌빈도 값의 곱으로 나타낸다. 다음 식 (2.1)은 문서의 우선순위  $W_{ij}$ 를 문서  $j$ 에 있는 단어  $i$ 의 가중치로 표현하고 있다.<sup>[20]</sup>

$$W_{ij} = freq_{ij} \times (\log \frac{N}{DF_i} + 1) \quad (2.1)$$

여기서  $freq_{ij}$ 는 단어  $i$ 의 문서  $j$ 에서의 출현빈도,  $N$ 은 총 표본 문서의 개수,  $DF_i$ 는 단어  $i$ 를 포함하는 문서의 개수이다. 역문헌빈도를 문서의 우선순위 결

정에 활용하는 이유는 모든 문서에 고르게 출현하는 단어의 경우 검색을 위한 색인어로서의 가치가 낮으므로 가중치를 낮추고자 함이다.

### 2.1.3 질의서버

질의서버는 사용자에게 질의어를 입력 받아 색인화를 참조하여 검색결과를 출력해주는 질의 처리 모듈이다. 이 모듈은 색인데이터베이스에서 주어진 질의어와 색인 정보를 패턴 비교하여 일치하는 관련 정보들의 목록을 결과로 제공한다. 질의 처리 모듈은 그림 2.4와 같이 입력 모듈, 검색 모듈, 순위 모듈, 출력 모듈로 구성된다. 입력 모듈은 사용자로부터 검색할 질의어를 입력받는 인터페이스 부분이다. 검색 모듈은 입력받은 질의어와 일치하는 색인어 정보를 색인데이터베이스에서 검색한다. 순위 모듈은 검색된 결과들을 이용하여 색인화 알고리즘에서 결정된 결과와는 다른 출력 형식을 사용자가 지정할 경우 새로운 기준에 따라 출력 순위를 재조정한다. 마지막으로, 출력 모듈은 한 페이지에 출력할 개수 만큼씩 검색 결과를 사용자에게 보여 준다.

질의 처리 모듈에서 집합적 해석이 필요한 연산은 부울 대수를 이용하여 색인데이터베이스를 검색한다. 집합적 해석은 하나 혹은 두 개의 질의어를 질의 집합에서 AND, OR, NOT의 논리 연산자를 사용하여 관계 연산을 실시하고 질의 결과를 구성한다. 부울 대수 논리를 이용한 질의는 정보 요구를 나타내는 용어인 색인어와 색인어들간의 관계자인 논리연산자로 구성되므로, 정보요구가 비교적 정확하고 연산자의 서술 과정이 인간의 논리적 연산과정과 유사하므로 사용이 편리하다는 장점을 가진다. 그러나 이러한 구조를 갖는 대부분의 범용검색엔진들은 인터넷 상의 정보가 방대함으로 인하여 여러 문제점들을 내포하고 있다.

첫째, 웹에 존재하는 정보 사이트들의 수가 매우 많고 계속 증가하고 있다는 점이다.

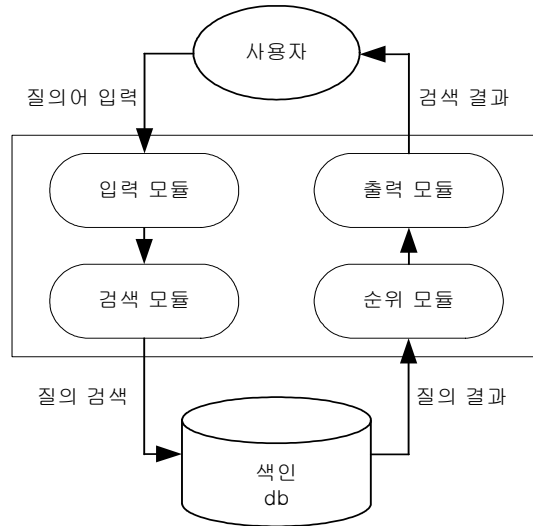


그림 2.4 질 의 처리 모듈

둘째, 질의어 입력을 통한 검색은 질의어의 선정 기준이 애매하여 사용자의 주관적 판단에 의존하는 경우가 대부분이다. 셋째, 검색된 결과가 아주 많은 경우 사용자에게 어떤 기준으로 정보를 여과해 줄 것인가 하는 점이다. 넷째, 사용자 수준에 맞는 적절한 검색 방법과 기준이 미비하다는 것이다. 다섯째, 인위적으로 사용자 수준에 맞는 검색을 하기 위해서는 질의어 자체를 계속 세분화해야 한다. 따라서 이러한 문제를 해결하기 위해서는 사용자의 성향이나 수준에 따라 정보를 필터링 하는 지능적인 에이전트의 개발이 필요하다.

## 2.2 지능형 에이전트의 고찰

지능형 에이전트란 특정한 목적을 위해 사용자를 대신해서 학습능력이나 추론 능력 및 계획 능력과 같은 지능적인 능력을 가진 자율적 프로세스이다. 여기서 사용자를 대신해서 작업한다는 것은 사용자의 의도를 파악해서 사용자의 요구사항을 처리해 준다는 것을 의미한다. 지능형 에이전트에 대한 개념은 연



구 집단마다 다른 정의가 가능하지만 다음과 같은 공통적인 특성을 지니고 있다.

- 자율성 : 사용자의 직접적인 간섭 없이 스스로 판단하고 동작하여 자신의 작동과 내부 상태에 대한 제어 수단을 갖는다.
- 사회성 : 에이전트 통신 언어를 사용하여 사용자와 다른 에이전트들과 상호 동작을 수행한다.
- 반응성 : 에이전트를 둘러싼 외부 환경을 인식하고 그 안에서 일어나는 변화와 상태를 파악하여 적절히 반응한다.
- 능동성 : 환경에 따라 단순히 반응하는 것이 아니라 주도권을 가지고 목표 지향적으로 행동한다.
- 시간 연속성 : 비연속적인 입력을 처리하고 종료하는 것이 아니라 후면에서 데몬(demon)과 같이 연속적으로 작업한다.
- 목표 지향성 : 주어진 문제를 해결하기 위해 최적의 수행 방법과 수행 순서를 스스로 결정한다.

지능형 에이전트는 기능과 역할에 따라 학습 에이전트, 인터페이스 에이전트, 데스크탑 에이전트, 모바일 에이전트, 전자상거래 에이전트, 인터넷 에이전트 등으로 구분된다. 또한 정보 검색과 관련된 인터넷 에이전트도 정보 가공 방법에 따라 정보검색 에이전트, 정보여과 에이전트, 정보통합 에이전트 등으로 나누어진다.

### 2.2.1 정보검색 에이전트의 지능적 학습 방법

현재 분산된 웹 환경에서 대부분의 정보검색 에이전트들은 사용자의 질의를 여러 웹서버들에게 동시에 브로드캐스트 하고 웹서버와 연동된 데이터베이스들에 의해 제공되는 결과를 사용자에게 웹 문서의 형태로 제시하는 무작위 추

출방법을 사용하고 있다. 그러나 이러한 방법은 불필요한 네트워크 자원의 접근으로 연관성 없는 정보를 결과로 반환하는 문제점을 지닌다. 그러므로 효율적인 분산정보검색을 위해서는 사용자가 원하는 문서를 선택적으로 검색하는 방법이 필요하다. 이러한 효율적 검색을 위해서는 기본적으로 지능적인 검색 방법을 필요로 한다. 에이전트가 지능적인 검색을 수행하기 위해서는 사용자에게 적응할 수 있는 적응성이 필요하다. 예를 들면, 웹을 탐색하는 에이전트는 현재 탐색을 의뢰하는 사용자가 누구인가에 따라서 다른 탐색 결과를 제공할 수 있어야 한다. 에이전트가 이러한 적응성을 갖지 못한다면 아무리 방대한 정보를 제공하거나 빠른 검색 속도를 제공한다고 하더라도 사용자가 만족하는 결과를 기대할 수 없다.

적응성을 가진 대부분의 에이전트는 사용자의 성향을 학습하기 위해 귀납적 기계학습방식을 이용하고 있지만 지식의 검색 방법과 표현 방법에 따라 여러 가지 지능적인 학습방법이 연구되고 있다. 지능적인 학습방법에는 분류(classification), 신경망(neural network) 등 대부분의 학습문제에서 문제가 ‘<입력값, 출력값>’쌍 집합으로 주어지고 주어진 입력값과 출력값을 만족하는 함수 F를 찾아내는 것이 목적인 감독 학습 방법<sup>[21]</sup>, 클러스터링이나 새로운 규칙발견 분야에서 주어진 입력값에 대한 출력값이 주어지지 않고 학습 목표가 주어진 문제에 아주 의존적으로 진행되는 무감독 학습 방법<sup>[22]</sup>, 웹 에이전트와 같은 동적인 환경에서 입력 값에 대한 결과가 피드백으로 학습 시스템에 영향을 미치면서 학습의 결과를 판단하는 강화 학습 방법<sup>[23]</sup>, 다수의 확률 변수들 간의 관계를 표현하는 그래프 모델로 이미 알고 있는 지식을 이용하여 조건부 확률을 계산하는 베이저안 이론(bayesian theorem)에 기반을 둔 확률 그래프 모델 방법<sup>[24]</sup>, 자연계의 진화 과정을 컴퓨터상에서 시뮬레이션 하여 복잡한 문제를 해결하는 방법으로 복제, 교차, 돌연변이 등의 연산자를 이용하여 최적해를 찾는 진화 학습 방법<sup>[25]</sup> 등이 있다.

그런데 정보검색 에이전트와 관련해서는 웹 페이지에 존재하는 단어들에 대한 관심도, 선호도를 고려하여 관련 지식을 제공하기 위해서 가중치를 가진 용어들을 벡터 형태로 표현하는 TF-IDF 알고리즘<sup>[26]</sup>과 사용자를 모델링 하여 프로파일을 작성하기 위해 주어진 자료의 속성 값을 트리의 최상위 노드로부터 최종 노드까지 정렬하여 입력 데이터의 카테고리를 분류하는 방식인 ID3<sup>[27]</sup>와 같은 알고리즘 등이 연구되고 있다. 또한 사용자들의 성향에 따라 환경에 적응하면서 사용자로 하여금 보다 효과적인 선택을 할 수 있도록 하기 위하여 위에서 기술한 방법들을 접목시키려는 연구도 활발히 진행되고 있다.

## 2.2.2 지능형 에이전트들의 특성

지능적 학습 방법들을 이용하여 개발된 웹 브라우징 에이전트 및 백그라운드 에이전트들의 종류별 특성을 살펴보면 다음과 같다.

### 1) Personal WebWatcher

카네기 멜론 대학에서 만든 Personal WebWatcher는 사용자와 웹 사이에 상주하는 인터페이스 에이전트로서 웹 브라우저 상에서 사용자의 행동을 모니터링 하여, 사용자의 적응력을 높인 에이전트이다. 이 시스템은 사용자의 관심도를 학습하기 위해 클라이언트 내부에 감시 프로세스가 백그라운드로 실행되면서, 사용자가 브라우저를 사용하는 패턴과 사용자의 행동을 관측하는 무감독 학습 방식을 이용한다.

이 방법은 제공된 문서의 관심도를 사용자에게 직접 확인하지 않고 확장 하이퍼링크 표현방법을 사용하여 사용자의 관심 문서들을 수집한다. 이렇게 모아진 관심 문서들을 통합하여 사용자별 프로파일을 만들고, 학습 단계에서 요청된 페이지들을 분석하여 사용자 프로파일을 갱신한다. 수집 및 갱신된 문서들은 사용자의 관심분야에 따라 문서 형태와 내용에 대한 개념 모델을 만들고 사용자가 정보 검색 시에 이 개념 모델을 이용하여 해당 문서에 대한 관심여

부를 예측한다.

Personal WebWatcher의 기능모듈은 그림 2.5와 같이 크게 세 부분으로 구성되어 있다. 먼저 사용자가 검색하는 웹 페이지를 모니터링 하는 프락시 모듈 부분으로 사용자가 현재 검색하고 있는 문서의 위치 및 내용을 프락시 저장영역에 보관하는 기능을 담당한다. 두 번째는 학습엔진 모듈 부분으로 모니터링 과정에서 수집된 결과를 분석하기 위하여 수집된 문서내용을 벡터 표현으로 변환, 불용어 제거, 어형변환 등을 통해 중요 단어를 추출하는 특징 추출을 수행한다.

사용자가 관심을 가지는 문서와, 그 문서에 대한 주요 단어를 사용자의 관심 영역으로 분류되는데 이를 위하여 TF-IDF, 베이지안 확률(bayesian probability)을 이용한 학습을 수행한다. 그리고 각 주제별 클래스 내의 중요 키워드를 추출하여 사용자의 개념 모델을 완성한다. 세 번째는 추천 모듈 부분으로 사용자가 웹 브라우저에서 검색을 수행할 때 각 문서에 대한 링크 중 사용자의 관심과 유사한 정보를 사용자에게 제안하고, 제안된 추천링크에 대한 사용자의 반응은 다시 모니터링 하기 위해 피드백 시킨다.

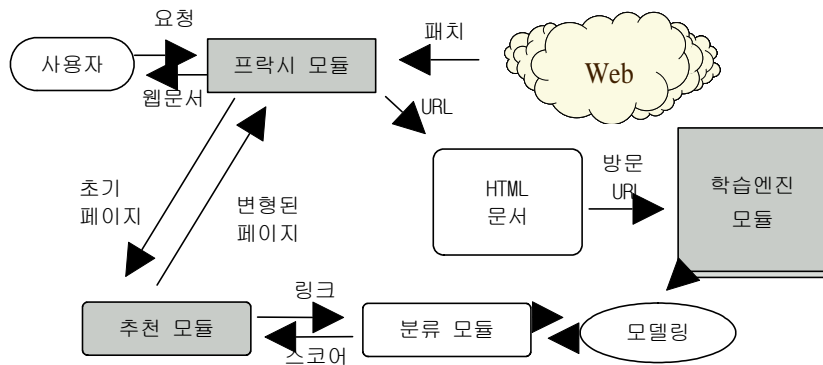


그림 2.5 Personal WebWatcher의 정보검색 모델

## 2) Letizia

Letizia는 MIT Media Lab에서 개발된 지능형 에이전트로서 사용자의 행위 정보를 수집하는 클라이언트형 에이전트이다. 사용자가 브라우저를 사용할 때 에이전트가 사용자의 행태를 추적하여 사용자의 현재 위치로부터 자동적으로 링크를 탐색하여 사용자의 관심분야를 HTML 문서로 제안한다.

범용검색엔진은 사용자가 질의한 내용에 대해 관련 웹사이트 정보들을 링크들로 구성된 웹 페이지로 반환 하지만, 이 에이전트는 사용자와 직접적인 상호작용 없이 사용자가 접근하는 URL을 프로파일로 기록하고 갱신하였다가 이를 근거로 사용자가 브라우징하는 동안에 병행적으로 웹을 검색하여 관심 웹사이트들을 브라우저로 보여준다.

이 에이전트가 사용자의 행위지식을 학습하고 연관성 높은 정보를 추천하는 과정은 그림 2.6에서처럼 다음과 같이 일곱 단계로 나누어 진행된다.

- 단계 1 : 패턴 관찰 모듈을 통해 사용자의 추천 URL을 기록한다.
- 단계 2 : 특징 추출 모듈을 통해 읽어들이는 페이지의 특징을 추출한다.
- 단계 3 : TF-IDF 기법을 통해 페이지를 분석하고 빈도를 측정한다.
- 단계 4 : 프로파일 발생 모듈로 발행한 프로파일을 저장한다.
- 단계 5 : 웹에서 추출된 문서를 문서 표현 모듈을 통해 기록한다.
- 단계 6 : 문서 분류 모듈을 통해 문서를 분류한다.
- 단계 7 : 분류된 문서를 사용자에게 추천한다.

이 에이전트는 사용자의 브라우징 행위로부터 사용자의 관심분야를 경험적 추론 방법의 일종인 최적우선탐색 방법으로 검색을 수행한다. 이 탐색 방법은 사용자가 관심분야에 대한 URL 링크 방문정보를 계속 추적하면서 방문 URL의 연결고리를 프로파일로 저장한다. 하이퍼링크가 계속해서 추적하고 저장되

면 사용자는 그 주제에 대해 관심이 있다는 것을 의미하고 사용자가 목표 문서를 저장하지 않거나 링크의 추적을 중간에서 그만두면 그 주제에 대해 사용자는 무관심한 것으로 간주한다. 같은 주제에 대해 반복적으로 방문하면 그 분야에 대한 관심도는 증가되어 다른 주제 보다 우선적으로 추천이 고려된다.

이러한 관심도 측정 전략은 정보 검색과 정보 필터링 전략을 통해 시스템의 유휴시간 동안 백그라운드에서 저장된 방문 패턴 정보를 대상으로 자동으로 연관성 검사를 수행하고 사용자의 요구가 있을 때 관련 문서에 대한 링크의 형태로 사용자의 다음 행동을 권고하게 된다. 이때 해당 방문 페이지에 머문 시간(elapsed time)도 연관성 정도에 대한 긍정적 혹은 부정적 증거를 구분하는 중요한 요소가 된다. 이 에이전트는 적응성에 기반한 인터페이스 에이전트이기 때문에 어떤 결정을 내리기 위해 필요한 참조 지식을 룰 형태로 지식베이스에 내장시켜 사용하는 것이 아니라 사용자의 구체적인 행위로부터 학습되어진 행위지식을 이용한다.

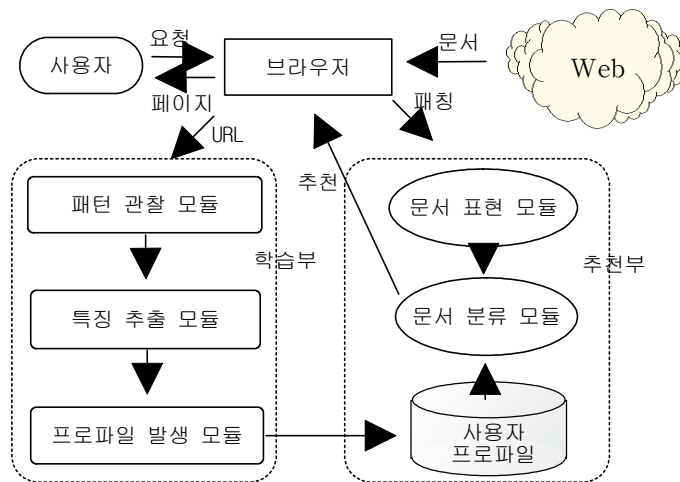


그림 2.6 Letizia의 정보검색 모델

### 3) WiseWire

WiseWire는 WiseWire사에서 만든 지능형 에이전트로서 정보 검색, 정보 검출, 정보 전달과정을 조정하고 추천하기 위해 지능적이고 경험적인 모델을 사용한다. 이 에이전트는 문서 내용과 사용자의 방문 기록을 분석하여 프로파일을 구성하고 축적된 정보를 이용하여 사용자들에게 각 개인이 원하는 정보를 검색할 수 있도록 지원한다. 프로파일에 축적되는 정보는 문서 내용과 사용자의 방문 패턴 정보 외에 사용자들의 직접적인 의견도 학습과정에서 피드백 정보로 저장된다.

웹 에이전트는 주제별 검색엔진에서 사용하는 일반적인 주제들을 사용자 검색에 제공한다. 각 주제는 그 하위 계층에 더욱 세분된 주제들을 포함하고 있다. 세분된 주제들 가운데 사용자가 관심을 가지는 주제를 선택하면, 개별 주제에 대한 관련 정보를 제공한다. 사용자는 제공된 정보들에 대해 관심 정도와 관심 분야에 따라 해당 문서와의 관련성을 평가받고 사용자의 선택과 평가를 기반으로 각 사용자의 문서에 대한 관심도가 학습된다. 이 시스템이 사용자의 관심도를 학습하기 위해 사용하는 속성은 문서의 내용, 관련성, 저자, 소스, 검색 날짜 등이다. 사용자는 관심정도를 평가하기 위해 0에서부터 10까지 세분화된 수치 값을 입력하고, 시스템은 이를 학습하여 사용자의 관심 분야를 개인화 시킨다. WiseWire 에이전트가 사용자의 관심도를 측정하고 학습하는 과정은 그림 2.7과 같이 학습 모듈을 통해 사용자가 방문한 문서의 속성 정보를 추출하여 문서 획득 모듈과 정보 추천 모듈에 피드백 정보로 제공된다. 이 정보들은 문서 정보와 추천 정보들과 함께 필터링 과정에서 정보 검출의 제약 정보로 다시 제공되어 사용자에게 보다 적용된 정보로 활용된다.

이 에이전트는 사용자 자신의 특성을 표현하는 속성 정보에 접근하여, 이미 생성된 모델의 각 항목에 대해 사용자가 학습을 직접적으로 지시할 수 있는 메커니즘을 제공하고 있다.

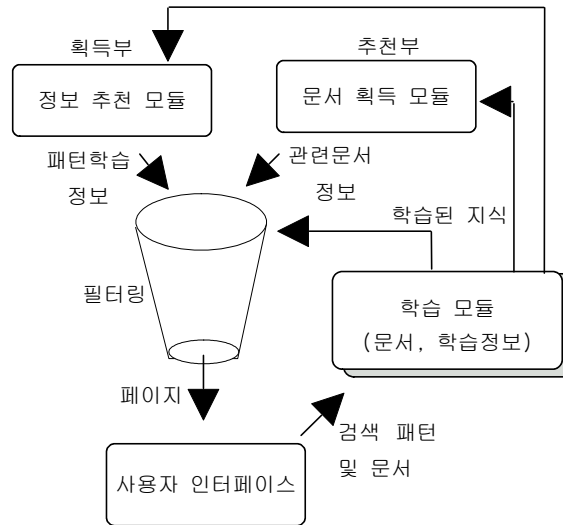


그림 2.7 WiseWire의 정보검색 모델

즉 학습할 주제나 문서에 대해 사용자가 검색에 필요하거나 불필요한 항목들을 필터링하기 위해 긍정적 항목(positive hint)과 부정적 항목(negative hint)을 부여하고 사용자 자신의 관심 분야에 대한 정보를 집중적으로 검색하거나 부정적 항목으로 부여한 질의어와 관련된 내용은 관심대상에서 제외시키는 작업이 가능하다. 이러한 과정을 통해 생성된 사용자의 특성 모델은 사용자별로 개인화 되고 적응력 있는 검색 기능을 제공할 수 있다.

또한, 이 시스템은 협력학습 기능을 수행한다. 협력학습 기능이란 비슷한 관심을 보이는 사용자들의 가상의 공동체를 구성하고, 그룹내의 개별 사용자들의 학습 결과를 공동체 내의 다른 사용자들의 관심 정보로 재학습하는 방법이다. 이와 같은 상호협력 필터링을 이용하면 공동체 내의 한 사용자가 학습한 내용을 공동체 내의 다른 사용자가 함께 이용할 수 있기 때문에 검색 효율을 높일 수 있는 이점이 된다.



#### 4) 기타 지능형 에이전트

이상의 대표적인 지능형 에이전트 시스템 외에도 SMART, SavvySearch, GLOSS(Glossary of Servers Server), Amalthea와 같은 정보 검색 에이전트 시스템이 있다.

먼저 SMART 시스템은 Salton Back에 의해 제안된 벡터공간모델 이론에 기반한 정보검색 에이전트이다. 이 시스템은 정보검색 연구 분야에 하나의 프레임워크로 인식되기 때문에 대부분의 검색분야의 학문적 연구에서 많이 참조되고 있다. 또한 색인화, 검색, 평가 부분은 하나의 표준적인 모델로 인식되며 제한적인 영역의 정보만을 사용자에게 제공할 수 있는 지능적인 기법도 포함하고 있다. 이 기법은 모든 문서의 중심용어 벡터를 포함하는 군집과일을 색인으로 사용하여 사용자 질의의 유사도를 측정하는 방법이다.<sup>[28]</sup>

GLOSS 시스템은 스탠포드 대학에서 개발한 정보검색 에이전트로서 사용자의 특정 질의를 만족시키는 정보를 제공하기 위해 정보저장소의 위치를 지시해 주는 메타데이터를 통해 사용자는 자신이 찾고자 하는 정보를 선택할 수 있게 한다. 메타데이터의 구성은 각 용어에 대한 문서빈도나 용어가중치의 합과 같은 통계치를 기반으로 사용자의 질의어와 데이터베이스 내의 각 문서 사이의 관련도를 계산하는 방법을 사용한다.<sup>[29]</sup>

SavvySearch 에이전트 시스템은 대표적인 메타 검색엔진으로 콜로라도 주립대에서 개발하여 서비스를 시작했으며, 현재는 CNET Inc.사에서 서비스를 제공하고 있다. 이 시스템은 사용자의 경험적 정보검색 결과를 이용하여 사용자의 질의어와 각 문서 사이의 관련도를 강화학습법으로 구하고 이를 바탕으로 임의의 질의에 대한 문서의 관련도를 계산하는 방법을 사용한다. 이러한 방법은 데이터베이스로부터 통계적인 정보를 제공받지 않고도 예제질의들에 대한 검색결과들을 통해서 학습을 수행하므로 분산된 환경에서 능동적 적응성을 갖는다.

Amalthaea 시스템은 MIT에서 개발한 멀티에이전트로 개인화된 필터링기능과 정보 사이트의 모니터링 기능이 학습에 의해 계속적으로 발전하는 멀티에이전트 시스템이다. 이 시스템은 사용자 개개인의 흥미모델에 기반해 사용자들이 희망하는 잠재적 사이트를 찾아준다. 시스템의 구성 요소는 웹을 탐색하고 웹에서 특정 정보를 찾아내는 정보검색에이전트(IRA)와 찾아낸 정보를 적절하게 필터링하는 정보필터에이전트(IFA)로 이루어진다.

이 시스템의 학습 전략은 두 개의 에이전트 그룹이 상호 협조하거나 경쟁하는 진화론에 기반을 두고 있다. 즉 오랫동안 사용되지 않는 에이전트는 도태되고 자주 사용되는 에이전트들끼리는 서로의 질의어 벡터를 계승받은 자손 에이전트를 새롭게 생성시켜 계속 생존된다. 이러한 과정의 반복으로 IRA와 IFA의 집단은 점차로 특정 사용자의 성향에 맞는 정보를 제공하는 방향으로 학습되어 간다.

표 2.3은 SMART, GLOSS, SavvySearch 시스템에서 데이터베이스인  $DB$ 가 주어졌을 때 용어  $t_i \in T$ 에 대한 문서 데이터베이스  $db_j \in DB$ 의 용어관련도  $\Psi_{db}(t_i)$ 와 주어진 질의  $q \in Q$ 에 대한 데이터베이스  $db_j \in DB$ 의 질의관련도  $\rho_{db_j}(q)$ 를 통해 관련도 평가 함수를 나타내고 있다.<sup>[30]</sup>

표 2.3 각 시스템별 관련도 평가함수

시스템 \ 관련도	용어관련도 $\Psi_{db}(t_i)$	질의관련도 $\rho_{db_j}(q)$
SMART	$\frac{ db_j(\{t_i\}) }{ D_{db_j} }$	$\frac{\sum_{t_i \in q} \Psi_{db_j}(t_i)}{\sqrt{ q } \cdot \sqrt{\sum_{t_i \in T} \Psi_{db_j}(t_i)^2}}$
GLOSS	$ db_j(\{t_i\}) $	$\frac{\prod_{t_i \in q} \Psi_{db_j}(t_i)}{ D_{db_j} ^{ q -1}}$
Savvy Search	$\frac{\sum_{q \in Q' \text{ such that } t_i \in q}    db_j(q)   }{ \{db \in DB \text{ such that } db(t_i) \geq 1\} }$	$\frac{\sum_{t_i \in q} \Psi_{db_j}(t_i)}{\sqrt{\sum_{t_i \in T} \Psi_{db_j}(t_i)}}$

이러한 유사도 평가 함수를 통해 각 시스템은 지능적인 검색기능을 부분적으로 제공하고 있으나 대부분의 정보 검색 에이전트에서는 사용자별 선호도를 고려할 때 매우 제한적인 가정에 의존하여 검색함으로써 낮은 관련성 수준의 결과만을 제공하고 있다. 예를 들어 SMART 시스템의 경우 대규모 문서들이 질의어별로 색인데이터베이스에 균일하게 분포되어 있는 경우에는 좋은 검색 결과를 가져 올 수 있지만 대부분 용어벡터공간에서 질의어들의 분포도를 사전에 예측하기는 어렵기 때문에 종종 잘못된 결과를 가져오기도 한다. 또한 GIOSS 시스템의 경우 검색 시에 주어지는 질의결과는 일반적인 통계정보에 근거한 사용자별 추천 기능일 뿐 사용자의 선호도나 관심분야를 학습하여 개인화된 정보로 제공하는 기능은 포함하고 있지 않다.

따라서 이러한 시스템들은 항상 정확한 색인데이터베이스나 프로파일 정보를 가지고 있어야 한다는 제약적 선행 조건으로 인해 주어진 가정을 만족하지 않는 환경에서는 효과적인 검색 결과를 제공하지 못하는 문제점을 지니고 있다. 또한 연관성이 높은 관련 정보들을 의미론적으로 분석하는 능력도 떨어져 개별 사용자의 의도나 특성에 맞는 결과를 제대로 제공해주지 못하고 있다.

특히 자연어 처리를 통해 문서 내에 있는 문장을 검색해 내는 검색 에이전트들은 자연어 자체의 모호성 때문에 문장의 의미 분석은 매우 어려우며 이러한 문제 때문에 일련의 복잡한 분석과정을 거치고도 통계적인 방법을 사용하는 검색엔진에 비해 향상된 성능을 보이지 못하는 경우가 종종 있다. 따라서 본 논문에서는 사용자의 경험적 지식을 검색에 활용할 수 있는 연관규칙 탐사와 사례기반 추론기법을 검색 에이전트에 적용하여 사용자의 관심분야에 따라 개인화된 서비스를 제공하고자 한다.

## 제 3 장 지능형 웹 검색을 위한 이론적 연구

### 3.1 KDD 관련 기술

연관규칙 탐사와 사례기반 추론 기법은 KDD(knowledge discovery in database)에서 지식획득과 발견을 위해 연구되어온 분야이다. KDD는 로컬 데이터베이스로부터 유용한 정보 및 지식을 발견하기 위해 데이터의 선택 및 정제, 보완, 변환, 마이닝 기법의 적용, 모형의 평가라는 과정을 거치는 데이터마이닝 기술이다. 이것은 로컬 데이터베이스 상에서 대량의 데이터에 내포되어 있는 데이터간의 의미 있는 상호관련성과 유용한 패턴을 추출한다. 상호관련성과 패턴의 추출은 다단계적인 반복과 상호작용적인 과정을 거친다.<sup>[31]</sup>

이러한 데이터마이닝 기술은 최근 웹을 통한 정보와 지식의 획득이 일반화되어 감에 따라 웹 데이터마이닝 기술로 점차 발전하고 있다. 웹에서 이용되는 데이터마이닝 학습 이론은 군집화, 순차패턴, 웹 방문패턴, 연관규칙 탐사, 사례기반 추론 등이 있다.<sup>[32]</sup>

먼저 군집화란 유사한 특성을 갖는 여러 객체를 몇 개의 그룹으로 클러스터링 하는 마이닝 기법으로 무감독 학습의 특징을 지니면서 정해지지 않은 다량의 데이터 집합에 대해 데이터베이스의 제한적인 요소들의 충족 정도가 평가기준이 된다. 웹 데이터마이닝에서 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 나뉘어 진다. 파티션 최적 알고리즘은 k개의 모든 가능한 파티션을 열거한 후 군집화가 얼마나 잘 형성되었는지 나타내는 척도로서 함수 값이 가장 좋은 것들로 그룹들을 정한다. 숫자 속성데이터를 군집화하는 가장 대표적인 파티션 최적 알고리즘으로 K-평균 군집화(K-means

clustering) 알고리즘이 있다. K-평균 군집화 알고리즘은 각 개체를 가장 가까운 군집의 중심점에 할당하는 방법으로 먼저 n개의 속성으로 구성된 군집의 초기 중심점 k개를 선택한다. 각 개체를 현재 개체와 가장 가까운 중심점을 갖는 군집에 할당한 후 군집의 새로운 중심점을 계산한다. 각 개체의 할당에 변화가 없을 때까지 위의 단계를 반복하여 최종적으로 k개의 새로운 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 거리로 나타내는데 거리 결정은 유클리드 거리, 맨해튼 거리, 클래스 간 평균거리, 클래스 내 평균거리 측도 방법이 있으며 계층적 클러스터링 최적 알고리즘의 군집방법에는 최단연결법, 최장연결법, 평균연결법, 중심연결법, 중위수연결법 등이 있다.<sup>[33]</sup>

순차패턴은 각 사용자들의 한 트랜잭션 안에서 발생하는 페이지간의 연관규칙에서 시간적 변이 개념이 추가된 것이다. 즉 연관규칙은 트랜잭션 안에서 어떤 페이지들 간의 상호 연관적 관계를 통해 관련성을 평가하는 반면 순차패턴은 트랜잭션 상호간의 관계를 평가하는 것이다. 각 사용자들의 순차적인 트랜잭션을 사용자 순차집합이라고 하고 특정 사용자의 순차가 사용자 순차 집합에 속해 있다면 그 사용자는 시 순차(time sequential)를 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.<sup>[34]</sup>

웹 방문패턴은 웹사이트에 존재하는 문서 페이지들의 구성 경로를 찾는 방

법을 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 빈도수를 의미한다.<sup>[35]</sup>

데이터마이닝 기법의 적용에 있어서 중요한 점은 각 기법들이 특정 문제 영역에만 적합성을 지닌다는 것이다. 예를 들어, 순차패턴은 시간적 변이를 갖는 트랜잭션 상호간의 연관적 관계를 찾는 문제에는 유용 하지만 검색 에이전트와 같이 단어와 문서 상호간의 연관적 관계를 평가하는 문제에는 적합하지 않다. 따라서 범용적으로 적용 가능한 데이터마이닝 기법은 없고 특정한 사안별로 적합한 알고리즘을 선택하는 것이 기술적인 문제이다.

따라서 본 논문에서는 위의 여러 기법 가운데 개인화된 정보 제공과 웹 에이전트의 학습을 위해 지지도 및 신뢰도의 평가와 유사도 평가가 다른 학습 기법에 비해 상대적으로 용이한 연관규칙 탐사와 사례기반 추론 기법의 적용 타당성 여부를 보다 심도 있게 고찰하고자 한다.

### 3.2 연관규칙 탐사 기법

연관규칙 탐사 기법은 항목집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 찾는 방법으로서 데이터간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고자 할 때 이용되는 기법이다. 연관규칙 탐사 기법에서는 데이터의 트랜잭션 로그(transaction log)로부터 데이터간의 연관성 정도를 측정하여 사용자의 요구에 대해 연관성이 높은 추가적인 요구들을 그룹화 하여 동시에 제공할 수 있도록 데이터 상호간에 연관성을 부여하고 있다.

연관규칙 탐사 기법은 그림 3.1과 같이 가설을 바탕으로 두 단계의 측정 과정으로 구성된다. 단계 1에서는 사용자가 미리 정의한 최소 지지도를 만족하는 데이터 항목조합들만 추출하고, 단계 2에서는 단계 1에서 얻은 데이터의

부분 집합에서 생성된 규칙 중 사용자가 정의한 최소 신뢰도를 만족하는 규칙들을 탐색하여 최종 규칙으로 정하게 된다. 연관규칙 탐사기법의 성능은 단계 1에서 결정되며 단계 1에서 추출한 빈발항목집합(large item sets)을 확인한 후에 연관규칙의 신뢰도는 단계 2에서 평가된다. 예를 들어  $I=\{i_1, i_2, \dots, i_k\}$ 를 항목집합이라고 하고 트랜잭션들로 이루어진 데이터 셋(data set)  $D$ 가 존재할 때 각 트랜잭션은 고유한 트랜잭션 번호가 부여된다.  $X \Rightarrow Y$  형식에서  $X \subset I, Y \subset I$ 이고  $X \cap Y = \emptyset$ 일 때 연관규칙은 지지도와 신뢰도를 바탕으로 트랜잭션 데이터 셋에서 각 항목간의 연관성을 찾는 것을 의미한다. 지지도는 전체 트랜잭션에 대한  $X$ 와  $Y$ 를 포함하는 트랜잭션 비율을 의미하고 신뢰도는  $X$ 를 포함하는 트랜잭션에 대한  $Y$ 를 포함하는 트랜잭션을 말한다.

□ **가설(hypothesis)**

- ItemSet  $I$ 의 부분집합  $X$ 에 대해 트랜잭션  $T$ 와  $X \subseteq T$ 의 관계이면  $T$ 가  $X$ 를 만족한다고 정의한다.
- ItemSet  $X$  ( $X \subseteq I$ )를 만족시키는 트랜잭션 수를  $|X|$ 로 정의하고 전체 트랜잭션의 수를  $N$ 으로 표기한다.
- $X, Y \subseteq I$ 에 대해  $X \cap Y = \emptyset$ 의 특성을 갖는다.

□ **측정(measure)**

- 단계 1 : support  
(연관규칙  $X \Rightarrow Y$ 에 대한 지지도)

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

- 단계 2 : confidence  
(연관규칙  $X \Rightarrow Y$ 에 대한 신뢰도)

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

그림 3.1 연관규칙을 위한 지지도 및 신뢰도

사용자가 정한 최소지지도를 만족하는 항목집합을 빈발항목집합이라 하고 최소신뢰도를 만족하는 빈발항목 집합이 있으면 유효한 연관규칙이 있다고 말한다.

### 3.2.1 빈발항목집합의 정의

연관규칙을 정의하기 위해 다음과 같은 집합을 가정한다.  $I = \{i_1, i_2, \dots, i_k\}$ 을 선택된 항목(literal)들의 집합으로,  $D$ 를 전체 트랜잭션들의 집합으로, 각 트랜잭션인  $T$ 는  $T \subseteq I$ 인 항목들의 집합으로 가정한다. 각 트랜잭션에서는 항목들의 크기는 고려하지 않으며, 각 항목은 그 항목의 빈도 여부만을 나타내는 이진 변수로 설정한다. 또한 카디널리티(cardinality)  $k = |X|$ 인 항목집합  $X$ 를  $k$ -항목집합이라 부르며  $k$ -항목집합은  $k$ 개의 항목들로 구성된다.

이때 트랜잭션  $T$ 가  $X$ 의 모든 항목들을 포함한다면 ( $X \subseteq T$ ),  $T$ 가 집합  $X$ 를 ‘지지한다’고 한다. 또한  $X$ 를 지지하는 트랜잭션의 집합을  $T(X)$ 로 표현하고  $X$ 의 지지도를 축약형 ‘supp( $X$ )’로 정의한다. 이것은  $X$ 를 지지하는  $D$ 에 있는 모든 트랜잭션들의 개수를 의미한다. 만일 주어진 최소지지도  $\text{supp}(\min)$ 에 대하여  $\text{supp}(X) \geq \text{supp}(\min)$ 이라면 ‘집합  $X$ 는 빈발하다’라고 정의한다.<sup>[36]</sup>

최소 지지도를 사용하는 이유는  $D$ 에 대하여 관심 있을 정도로 빈발하게 나타나는 항목만을 대상으로 고려하기 위함이다. 빈발하지 않은 항목집합 즉 최소지지도를 만족하지 않는 항목집합들은 고려 대상에서 제외한다.

다음은 연관규칙탐사 알고리즘의 기초를 이루는 빈발항목집합들의 세 가지 규칙이다.

#### ■ rule 1(부분집합의 지지도)

만일 항목집합  $A, B$ 에 대하여  $A \subseteq B$ 이면,  $B$ 를 지지하는  $D$ 의 모든 트랜잭션들이 필연적으로  $A$  또한 지지하므로  $\text{supp}(A) \geq \text{supp}(B)$ 이다.



■ rule 2(빈발하지 않은 집합들의 상위집합)

만일 항목집합 A가 D에서 최소 지지도에 미치지 못한다면, 즉  $\text{supp}(A) < \text{supp}(\min)$ 의 룰(rule 1)에 의하여  $\text{supp}(B) \leq \text{supp}(A) < \text{supp}(\min)$ 이기 때문에 A의 모든 상위집합 B는 빈발하지 않을 것이다.

■ rule 3(빈발항목 집합들의 부분집합)

항목집합 B가 D에서 빈발하다면, 즉  $\text{supp}(B) \geq \text{supp}(\min)$ 인 룰(rule 1)에 의하여  $\text{supp}(A) \geq \text{supp}(B) \geq \text{supp}(\min)$ 이므로 B의 모든 부분집합 A는 D에서 또한 빈발 할 것이다. 만약  $A = \{i_1, i_2, \dots, i_k\}$ 가 빈발하면, 그것의 모든 k개의 (k-1)-부분집합들도 빈발하다. 그 역은 성립하지 않는다.

### 3.2.2 연관규칙의 정의

항목들의 집합인 X, Y가 I의 부분집합( $X, Y \subseteq I$ )이고 X와 Y는 서로 같은 원소를 갖지 않는 항목집합( $X \cap Y = \emptyset$ )이며 Y가 공집합( $Y \neq \emptyset$ )이 아니면 연관규칙은 “R: X→Y”형식으로 정의한다. 이때 X를 규칙의 조건부라 하고 Y를 결론부라 한다.  $X \cap Y = \emptyset$ 인 것은  $R: X \rightarrow X \cup Y$ 이고  $R: X \rightarrow Y$ 이기 때문에  $R: X \rightarrow X$ 와 동일한 의미를 갖는다. 만일 한 트랜잭션이 X를 지지한다면 또한 어떤 확률에 의해 Y도 지지할 것이라는 예측으로 이해될 수 있다. 이런 확률을 이 규칙의 신뢰도라 하고  $\text{conf}(R)$ 로 표시한다.

R의 신뢰도는 X를 지지하는 T에 대하여 Y 또한 지지할 조건부 확률로 정의되며 식 (3.1)과 같이 표시된다.<sup>[37]</sup>

$$\text{conf}(R) = P(Y \subseteq T | X \subseteq T) = \frac{P(Y \subseteq T \wedge X \subseteq T)}{P(X \subseteq T)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \quad (3.1)$$

D에 있는 규칙 R에 대한 지지도는  $\text{supp}(X \cup Y)$ 로 정의한다. 규칙의 지지도

는 얼마나 자주 적용되었는지를 나타내는 반면 신뢰도는 그 규칙이 얼마나 믿을 만 한지를 나타낸다. 규칙이 대상영역 D에서 유효(valid)하려면 충분한 지지도와 신뢰도를 가져야 한다. 그러므로 우리는 어떤 주어진 최소 신뢰도  $\text{conf}(\min)$ 과 최소지지도  $\text{supp}(\min)$ 에 대하여  $\text{conf}(R) \geq \text{conf}(\min)$ 이고  $\text{supp}(R) \geq \text{supp}(\min)$ 하면 규칙 R은 영역 D에 대하여 연관규칙이 성립된다. 즉 이 규칙이 성립되기 위한 필요조건으로서 연관규칙의 조건부와 결론부는 모두 빈발해야 한다. 이것은 하나의 데이터 오브젝트가 최소 지지도를 만족한다 하더라도 최소 신뢰도를 만족하지 않던가 최소 신뢰도를 만족하더라도 최소지지도를 만족하지 않는다면 이 규칙은 강한 상관관계를 표현할 수 없어 유효하지 않다. 연관규칙과 관련하여 rule 4, rule 5, rule 6가 있다.

■ rule 4(규칙의 합성)

만일  $X \rightarrow Z$ 이고  $Y \rightarrow Z$ 이면  $XUY \rightarrow Z$ 가 반드시 참인 것은 아니다. 그 규칙들이 X 또는 Y만을 지지하는 경우에만  $X \cap Y = \emptyset$ 이고 D에 있는 트랜잭션들이 Z를 지지하는 경우 집합 XUY는 지지도 0을 가지므로  $XUY \rightarrow Z$ 는 신뢰도를 갖지 않는다. 유사한 논리가 동일한 조건부를 가지고 규칙들의 합성에 적용된다. 즉  $X \rightarrow Y \wedge X \rightarrow Z \not\Rightarrow X \rightarrow YUZ$  이다.

■ rule 5(규칙의 분해)

만일 D에 대하여  $XUY \rightarrow Z$ 가 참이면  $X \rightarrow Z$ 와  $Y \rightarrow Z$ 가 반드시 참인 것은 아니다. 이런 것은  $\text{supp}(XUY) = \text{supp}(Z)$ 인 경우, 즉 예를 들어 Z가 한 트랜잭션에 있고 X와 Y가 동시에 같은 트랜잭션에 있는 경우에만 해당된다. 만일 X와 Y에 대한 지지도가  $\text{supp}(XUY)$ 보다 충분히 클 때, 두 개의 규칙들은 요구되는 신뢰도를 만족하지 못한다. 그러나 역으로,  $X \rightarrow YUZ \Rightarrow X \rightarrow Y \wedge X \rightarrow Z$ 는  $\text{supp}(XY) \geq \text{supp}(XYZ)$ 이고  $\text{supp}(XZ) \geq \text{supp}(XYZ)$ 이므로 성립된다. 그러므로 보다 적은 규칙들의 지지도

와 신뢰도는 원래의 규칙에 비교하여 더 증가한다.

■ rule 6(비이행성)

만일  $X \rightarrow Y$ 이고  $Y \rightarrow Z$ 이면  $X \rightarrow Z$ 를 추론할 수 없다. 예를 들어  $T(X) \subset T(Y) \subset T(Z)$ 이고 최소신뢰도가  $\text{conf}(\min)$ 이라고 할 때  $\text{conf}(X \rightarrow Y) = \text{conf}(Y \rightarrow Z) = \text{conf}(\min)$ 이라 하자. 상대적인 지지도에 기초하여  $\text{conf}(\min) < 1$  이기 때문에  $\text{conf}(X \rightarrow Z) = \text{conf}'(\min) < \text{conf}(\min)$ 인 관계를 만족하지 않으므로 이행적인 규칙을 얻을 수 없다.

### 3.2.3 연관규칙 탐사의 접근 방식

연관규칙에 대한 규칙탐사 알고리즘을 위해 주어진 탐색영역에서 탐사되는 연관규칙이 성립되기 위해서는 정의된 최소지지도와 최소신뢰도 이상의 값을 가져야 하므로 다음의 두 단계로 구성된다.

- 단계 1 : 빈발항목집합을 찾아낸다. 미리 결정된 최소지지도 이상의 트랜잭션 지지도를 가지는 항목집합들의 모든 집합들을 빈발항목집합이라 한다. 나머지 항목집합들을 작은 항목집합 (small item sets)이라 한다.
- 단계 2 : 탐색영역으로부터 연관규칙을 생성하기 위해 빈발항목집합을 사용 한다. 모든 빈발항목집합  $L$ 에 대해서  $L$ 의 공집합이 아닌 모든 부분 집합들을 찾는다. 각각의 부분집합  $A$ 에 대하여  $\text{supp}(A)$ 에 대한  $\text{supp}(L)$ 의 비율이 적어도 최소 신뢰도  $\text{conf}(\min)$ 이상이면 ( $\text{supp}(L)/\text{supp}(A) \geq \text{conf}(\min)$ ),  $A \Rightarrow (L-A)$ 의 형태의 규칙을 생성한다. 이 규칙의 지지도는  $\text{supp}(L)$ 이고 신뢰도는  $\text{supp}(L)/\text{supp}(A)$ 이다.

연관규칙의 성능은 빈발항목을 선정하는 단계에서 결정되며 신뢰성 측정인 유효성 검증을 통해 대상 후보 규칙의 수를 줄일 수 있다.

#### 1) 빈발항목집합 선정 방법

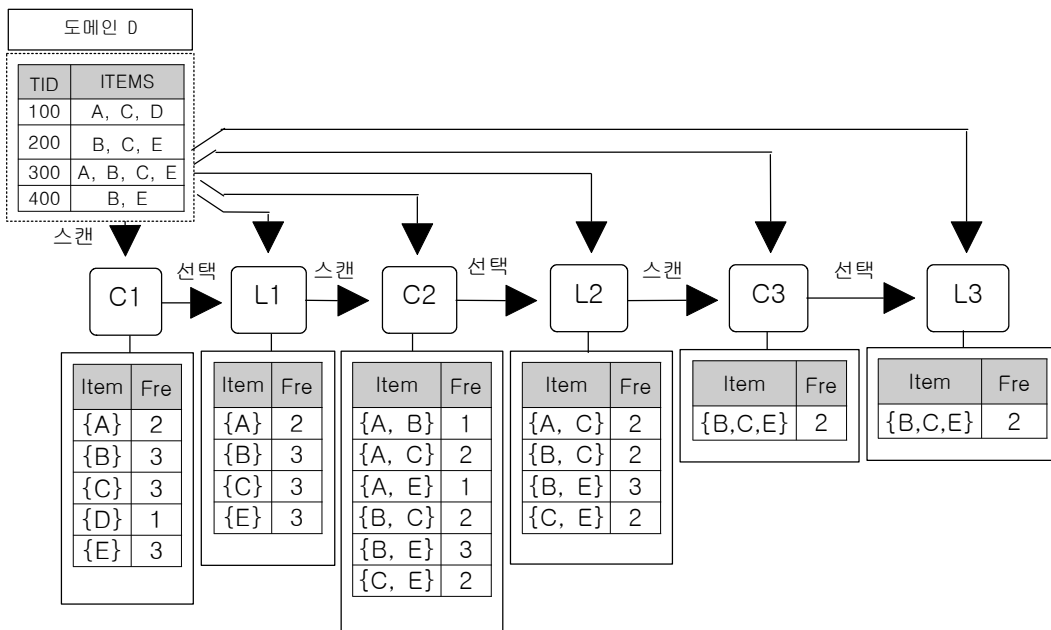
빈발 항목집합들의 수는 모든 대상 항목들의 멱집합(power set)으로 표현된다. 모든 항목들의 빈발 여부를 판단하기 위해 그 멱집합에 속한 모든 집합소들을 검증하면 항목들의 크기는 기하급수적으로 증가되고 결국 고갈(exhaust)탐색을 수행할 가능성이 있다. 이러한 소모적 탐색을 피하기 위해 발견적 알고리즘이 필요하다. 이 알고리즘은 후보라고 하는 빈발 가능성이 높은 항목집합들을 생성하고 이들 후보항목집합들 중에 실제로 빈발한 항목들을 찾기 위해 각 후보 항목집합들에 대한 지지도를 이용해 대상영역을 좁혀가면서 부분집합을 생성하는 과정을 반복적으로 수행한다. 후보 항목집합의 발생빈도를 계산하는 것은 상당량의 프로세싱 시간과 메모리를 요구하기 때문에 연관규칙 탐사 알고리즘의 성능은 후보들의 수에 비례한다.

빈발항목집합의 선정 과정은 대상영역에서 후보항목집합의 생성 단계와 빈발항목집합의 검색 단계를 거치는데 전형적인 방법론으로 Apriori 알고리즘이 있다. Apriori 알고리즘에서는 각 단계마다 빈발항목집합의 후보 집합을 구성하고 난 후에 각 후보 항목집합의 발생빈도 수를 계산하여 최소지지도를 만족하는 항목집합들을 빈발항목집합으로 선정한다. 그림 3.2는 대상영역인 도메인을 이용하여 빈발항목집합을 선정하는 과정을 보여주고 있다.

첫 번째 단계에서 각 항목의 발생 빈도수를 계산하기 위해 단순히 모든 트랜잭션들을 스캔하여 읽는다. 대상영역에서 항목별로 발생빈도를 계산하여 후보1-항목집합 C1을 결정한다. 최소 지지도가 2라고 가정하면 필요로 하는 최소지지도를 갖는 후보1-항목집합들의 항목으로 구성되는 빈발1-항목집합 L1을 결정한다. 빈발2-항목집합은 후보2-항목집합의 모든 부분집합도 역시 최소

지지도를 만족해야 하고 후보항목집합 C2의 생성은 이전 단계의 빈발항목집합인 L1\*L1을 사용하여 구성된다. 여기서 \*는 집합연산자이며, C2는  $\binom{|L1|}{2}$  개의 2-항목집합들로 이루어진다. |L1|이 크면  $\binom{|L1|}{2}$ 는 아주 큰 값이 될 수 있다. 다음으로 도메인 D에 속한 네 개의 트랜잭션들이 스캔되어 C2에 속한 각 후보 항목집합의 지지도가 계산된다.

후보항목집합들의 집합 C3은 L2에서 첫 항목이 같은 두 개의 빈발2-항목집합들을 먼저 확인한다. 예를 들면 {B, C}와 {B, E}에서는 B가 동일한 항목이다. 다음으로 Apriori 알고리즘은 {B, C}와 {B, E}의 두 번째 항목들로 구성된 2-항목집합 {C, E}가 빈발2-항목집합들에 속하는 지를 검사한다.



(Lk: Set of Large k-itemsets, Ck:Set of Candidate k-itemsets)

그림 3.2 빈발항목 선정과 지지도계산

2-항목집합  $\{C, E\}$ 가  $L_2$ 의 원소로 빈발항목이므로,  $\{B, C, E\}$ 의 모든 부분 집합들은 ‘빈발하다’고 결정하고  $\{B, C, E\}$ 는 후보3-항목집합이 된다.  $L_2$ 에서는 더 이상의 다른 후보3-항목집합을 구할 수 없다.

따라서 Apriori 알고리즘은 모든 트랜잭션들을 스캔하면서 빈발3-항목집합을 구성한다.  $C_3$ 을 기본으로 하여 도메인  $D$ 를 스캔하여  $L_3$ 을 찾아낸다.  $L_3$ 에서부터 구성될 수 있는 후보4-항목집합이 없으므로 여기서 빈발항목집합을 발견하는 과정은 종료된다.

## 2) Apriori 알고리즘에서 후보 항목집합의 생성과 지지도 계산

연관규칙의 문제는 AIS라는 알고리즘에서 처음 제기되었으며 대부분의 연관규칙탐사에 관한 연구는 빈발항목집합 생성과 규칙 형성을 분리하여 문제 해결을 시도하고 있다. AIS에서 많은 수의 후보집합 생성 문제는 Apriori-gen이라는 후보 항목집합의 생성전략을 통해 후보항목집합의 수를 줄일 수 있다. 이는 Apriori 알고리즘의 부함수로 대부분의 연관규칙 알고리즘에서 이용되고 있다. 이 방법은  $k$ -항목집합 중 빈발한 후보  $(k+1)$ -항목집합을 선택하기 위해 조인(join)단계와 전지(prune)단계를 반복적으로 적용하는 행위를 통해 후보항목집합의 수를 축소시켜 나간다. 즉 다음의 Apriori-gen() 함수에서 처럼 조인 단계는 빈발 항목집합  $L_{k-1}$ 를 입력으로 사용하고 그 집합에서  $k-1$ 개의 공통된 항목들과 다른 항목들을 결합하여 후보  $(k)$ -항목집합을 형성한다.

### **Algorithm Apriori-gen( $L_{k-1}$ )**

{조인단계}

insert into  $C_k$

select p.item<sub>1</sub>, p.item<sub>2</sub>, ... , p.item<sub>k-1</sub>, q.item<sub>k-1</sub>

from  $L_{k-1}$  p,  $L_{k-1}$  q

where  $p.item_1 = q.item_1, \dots, p.item_{k-1} < q.item_{k-1}$ ;

{전지단계}

forall itemsets  $c \in C_k$  do

forall  $k$ -subsets  $s$  of  $c$  do

if( $s \notin L_{k-1}$ ) then delete  $c$  from  $C_k$ ;

두 번째 집합의 가장 큰 항목이 첫 번째 집합의 가장 큰 항목보다 크다는 것을 제약 조건으로 사용하여 중복을 방지한다. 전지단계는 두 번째 단계에서 만들어진 후보  $(k)$ -항목집합의 부분집합  $(k-1)$ -항목집합들이 이미  $L_k$  안에 있는지를 검사하여 없으면 이 후보항목집합을 버린다. 예를 들어  $\{1, 3, 4, 6\}$ 과  $\{1, 3, 4, 8\}$ 이 빈발 하다면 두 항목집합은 조인되어 후보항목집합  $\{1, 3, 4, 6, 8\}$ 을 생성한다. 이 집합의 부분집합은  $\{3, 4, 6, 8\}$ ,  $\{1, 4, 6, 8\}$ ,  $\{1, 3, 6, 8\}$ 이며 만일 이들이 빈발하지 않거나 유효하지 않으면 버린다.

다음은 Apriori-gen 알고리즘에서 빈발항목집합인  $L_k$ 로부터 후보항목집합  $C_k$ 를 생성하는 과정을 예시한다.

- 단계 1 : 조인단계

$L_3 = \{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{a, c, e\}, \{b, c, d\}$

candidate 4-itemset =  $\{\{a, b, c, d\}, \{a, c, d, e\}\}$

- 단계 2 : 전지단계 :

3-subset of  $\{a, b, c, d\} = \{\{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{b, c, d\}\}$

3-subset of  $\{a, c, d, e\} = \{\{a, c, d\}, \{a, c, e\}, \{a, d, e\}, \{c, d, e\}\}$

each  $\{a, d, e\}, \{c, d, e\} \notin L_3 \therefore$  subset  $\{a, c, d, e\}$  is pruning

$\Rightarrow C_4 = \{\{a, b, c, d\}\}$

Apriori 알고리즘은 Apriori-gen을 사용한 첫 번째 알고리즘으로 후보 항목

집합들을 생성하는 단계와 지지도를 계산하는 단계로 구성되어 있다. 즉 각각의 패스에서 주어진 크기의 모든 후보들을 생성하기 위해 Apriori-gen을 호출하고 전체 대상영역을 스캔하여 각 후보들에 대한 지지도를 계산한다.

이 알고리즘은 k번째 패스의 계산 단계에 있는 트랜잭션 T를 읽을 때 트랜잭션 T에 의해 지지되는 모든 k-후보를 결정하고 그 후보들과 관련된 지지도 계수 값을 하나 증가시킨다. 다음은 Apriori 알고리즘을 나타낸다.

### Algorithm Apriori

```

L1={large 1-itemsets]
for(k=2;Lk-1≠∅;k++) do begin
    Ck=apriori-gen(Lk-1);           // New candidate
    forall transaction t∈D do begin
        Ct = subset(Ck, t);       // Candidates contained in t
        forall candidates c∈Ct do
            c.count ++;
    end
    Lk={c∈Ck|c.ount > supp(min)}
end

Answer =  $\bigcup_k L_k$ 

```

Apriori 알고리즘의 한가지 문제점은 지지도를 계산하는 과정에서 항목들 간의 카티전 프로덕트(cartesian product)를 통해 항목 후보 집합들이 폭증하여 수행 속도를 지연시키는 문제가 발생할 수 있다. 이 문제를 해결하기 위해 해쉬 트리(hash tree)에 후보 항목집합들을 저장함으로써 수행 속도를 개선시키는 방법이 제안되고 있다.



### 3.3 사례기반 추론

사례 기반 추론은 주어진 문제를 해결하기 위해 과거의 유사한 사례를 문제의 상황에 맞게 응용하여 해를 찾아가는 기법으로 새로운 요구에 대응하는 과거의 해를 채택하거나, 과거의 사례를 이용하여 새로운 상황을 설명하거나, 과거의 사례로 새로운 해를 평가하거나, 또는 새로운 상황을 이해하기 위해 사례로부터 주어진 문제에 대한 적절한 해를 추정하는 작업을 수행한다.<sup>[38]</sup>

사례 기반 추론은 Schank에 의해 제안된 동적 기억이론(dynamic memory theory)과 문제해결 및 학습에서의 회상(reminding)의 역할에 관한 연구를 기점으로 시작되었다. 동적 기억이론은 개별적인 경험이 어떻게 기억 속에 저장될 수 있는지, 각 경험들이 어떻게 결합되고 추상화 될 수 있는지 또 필요할 때 어떻게 검색되고 사용될 수 있는지를 기술하기 위한 골격을 제공한다.<sup>[39]</sup>

사례 기반 추론은 주어진 문제를 해결하기 위해 필요한 정형적 규칙을 찾기 힘든 문제 영역에 적용하는 것이 유용하며, 특히 과거의 경험으로부터 의사결정을 이끌어낼 수 있는 경우에 매우 효과적인 문제해결 방법이다. 과거에는 주어진 문제를 해결하기 위해 전문가시스템의 규칙기반 추론 기법(rule based reasoning)이 많이 사용되었다. 규칙기반 추론은 특정한 영역의 문제해결을 위해 전문가의 경험적 지식을 추출하여 생성규칙(production rule) 형식으로 정형화하고 전향추론이나 후향추론을 통해 해를 발견한다. 그러나 문제를 해결하기 위해 필요한 모든 지식을 미리 추출하여 정형화하는 것은 사실상 어렵고 문제 해결을 위해 매번 관련된 규칙을 순서대로 추적하는 것도 추론 시간을 지연시키는 큰 문제점 이었다. 이것은 사례 기반 추론이 규칙 기반 추론보다 단순하며 실현성이 높고 복잡한 문제영역에서 좋은 접근법이라 할 수 있다. 또한 사례 기반 추론은 유사 해를 찾을 때 과거의 문제와 현재의 문제간의 차이를 고려하여 이전의 해결책들을 현재의 문제에 맞게끔 변형하는 과정이 필요하다. 이 과정을 통해 지능시스템 구축의 가장 어려운 과제인 지식획득 문

제를 자연스럽게 해결할 수 있게 하고 귀납적 기계학습의 기반을 제공한다. 따라서 사례기반 추론 기법은 문제가 복잡하고 해를 구하는데 많은 시간이 요구되는 학습 영역에서 쉽게 해를 구할 수 있으므로 에이전트와 같이 다양한 사용자별 특성이 고려되는 지능적 시스템에서 효과적인 방법론이라 할 수 있다.<sup>[40]</sup>

### 3.3.1 사례의 구성과 모형의 표현

#### 1) 사례의 구성

사례기반 추론에서 사례의 구성은 추론에 영향을 미치는 중요한 요인 중 하나이다. 일반적으로 사례의 표현을 위한 구성항목의 선정에서는 구성항목의 기능성과 정보 획득의 용이성이 고려되어야 한다. 사례에는 문제의 내용과 해의 내용이 포함되며, 문제의 내용에는 문제의 요구사항이 서술되고 해의 내용에는 문제의 해결방법이 서술된다. 또한 사례는 보통 특정한 결과를 도출하는 주요한 특성의 리스트로서 표현된다. 하나의 사례를 표현하기 위해 필요이상의 특성이 사용되면 사례베이스의 크기가 너무 커져서 추론 효율이 저하될 수 있다.

사례 기반 추론 모형에서 사례는 프레임(frame), 객체(object), 의미망 등으로 표현될 수 있지만 대부분 프레임 표현을 많이 사용한다. 프레임 표현은 상속과 추상화 등 객체지향 개념의 특성을 지니고 있으며 사례베이스 구성을 관계형 자료구조로 표현하는 것이 가능하므로 활용성을 높일 수 있는 장점을 가지고 있다. 프레임은 스키마와 슬롯으로 구성되는 객체이므로 사례를 프레임으로 표현할 때 그림 3.3과 같이 사례는 스키마로 사례의 특성은 슬롯에 대응시킨다.

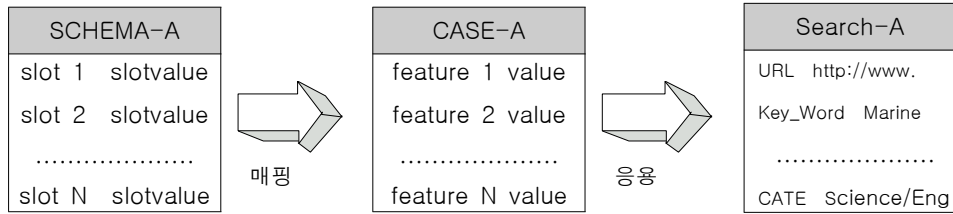


그림 3.3 사례의 프레임 표현 과정

## 2) 사례기반 추론 모형의 표현

사례기반 추론 모형의 개념적 모델은 표 3.1과 같이 주 영역과 보조영역으로 구성되어 있다. 주 영역은 그림 3.4와 같이 재귀적 반복작업을 수행하는 4R(Retrieve ⇒ Reuse ⇒ Revise ⇒ Retain)이라는 4개의 모듈로 구성되고 보조영역은 검색을 지원하기 위한 1개의 색인 모듈로 구성된다. 4R에는 사례베이스를 검색하는 회수모듈, 주어진 문제를 과거사례로부터 해결하고자 하는 재사용모듈, 필요한 경우 유사 사례로부터 해를 적용하기 위한 교정모듈, 그리고 교정된 해를 새로운 사례로 저장하는 보유모듈이 있다.

먼저 회수모듈은 과거의 사례로부터 주어진 문제의 해를 찾기 위해 사례베이스를 검색하는 서버모듈이며 사례기반 추론의 핵심 부분이다. 재사용모듈은 검색된 해를 재사용하기 위해 검색된 사례와 새로운 문제를 비교 분석하고 일치하는 사례가 발견되면 적합한 해로 제안한다.

표 3.1 사례기반 추론의 하부 항목

분야	모듈	문제의 역할
주 영역	회수	과거의 사례 검색
	재사용	불완전한 문제의 분석 및 이해
	교정	과거의 해를 주어진 문제로 적용
	보유	새로운 사례 저장
보조영역	색인	사례의 효과적인 저장 및 검색

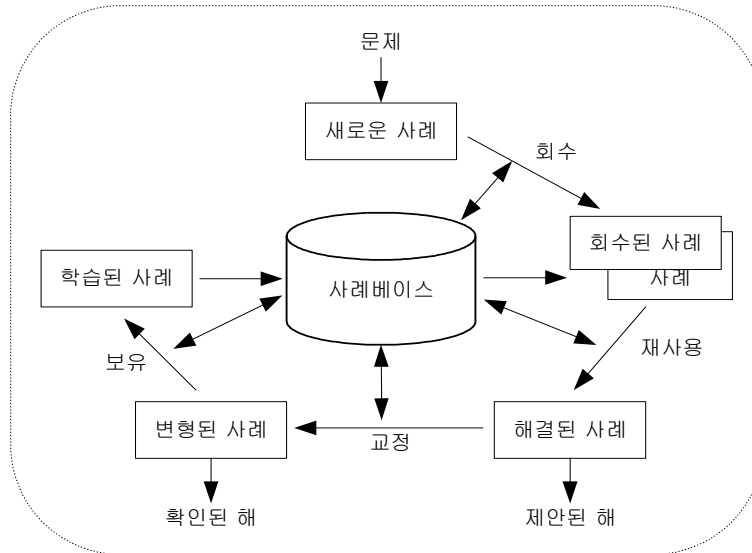


그림 3.4 사례기반 추론 모형의 개념적 표현

교정모듈은 적합한 해가 없는 경우 검색된 사례를 주어진 문제에 적용하기 위해 유사한 해를 도출하고 이를 평가하여 새로운 사례로 변형하는 기능을 담당한다. 교정모듈은 사례 기반 추론에서 가장 중요한 부분이나 구현하기 어렵고 새로운 문제 해결을 위해 주어진 문제마다 교정 과정을 거쳐야 하는 것은 사례기반 추론 시스템의 문제점이 될 수 있다. 교정 과정을 통과한 해는 주어진 문제에 다시 적용하는 시험 단계를 거쳐 성공 혹은 실패로 그 결과가 나타난다. 보유모듈은 제안된 해가 문제해결에 적합하면 현재 문제에 대한 데이터를 새로운 사례로 사례베이스에 저장하고 적합하지 않으면 실패의 원인을 확인하고 다시 교정 단계로 보내는 기능을 담당한다.

보조영역의 색인모듈은 사례의 저장방법을 결정하는 모듈로서 색인 방법에 따라 검색 효율에 큰 영향을 미친다. 일반적으로 색인화는 키워드별 연결관계를 개념그래프로 구성하는 방법을 사용하고 있다.

### 3.3.2 사례기반 추론 과정

사례 기반 추론의 세부 추론 절차는 그림 3.5에 나타난 것과 같이 사례 표현 단계, 유사사례 추출 단계, 근접해(ballpark solution) 제시 단계 및 적용 단계, 평가 및 수정 단계(evaluative reasoning), 사례베이스 보완 단계(update case-base)를 거친다.<sup>[40]</sup>

- 단계 1 : 사례 표현

일반적으로 사례는 경험을 표현하기 위한 방법이며 사례에는 과거 경험적 히스토리들이 내용으로 구성된다. 전형적으로 사례는 문제상황을 설명하기 위한 문제설명영역과 그 문제를 해결하기 위해 사용된 해를 표현하는 솔루션영역 그리고 솔루션이 결정되었을 때의 상태를 표현하는 결과영역으로 구성된다.

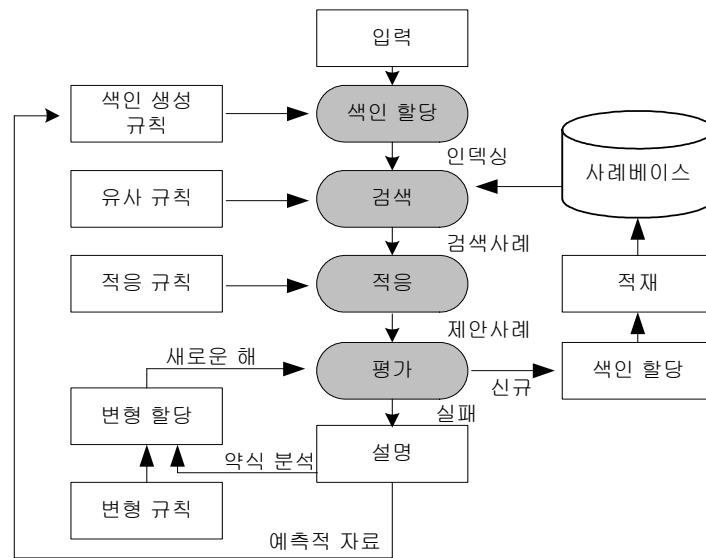


그림 3.5 사례기반 추론 흐름도

- 단계 2 : 유사사례 추출

이 단계는 새로운 질의어가 주어졌을 때 사례베이스로부터 가장 적절한 사례들을 추출하여 유사사례로 확정짓는 단계로서, 저장된 사례베이스의 색인화와 질의어를 패턴 비교하는 검색과정을 수행한다. 패턴 비교에서 정확하게 일치하는 자료가 없을 경우에는 유사도 평가라는 과정을 동반하게 되는데 유사도 평가 방법에 따라 질의어 검색 정보의 유효성이 결정된다. 유사한 사례를 검색하기 위해서는 사례가 부합하는지를 측정해서 가중치를 부여한 후, 사례를 순서화하는 것이 필요하다. 사례의 검색방법 중 대표적인 것으로는 최근접 이웃 검색(the nearest neighbor retrieval)과 귀납적 검색(inductive retrieval)이 있다. 귀납적 검색은 검색하기 전에 먼저 색인화하는 선행 작업이 필요하나 처리시간이 빠르고 최근접 이웃 검색은 모든 항목에 대해 상대거리를 측정하기 때문에 처리시간이 많이 소요되나 간단하기 때문에 적용이 용이하다.

- 단계 3 : 근접 해 제시 및 교정

이 단계는 색인화에 의해 유사사례가 근접해로 주어지며 후보해로 제공된 유사사례와 질의어가 정확하게 일치하지 않을 경우 제시된 후보 해를 새로운 상황에 맞게 적용할 필요가 있는데 이 과정을 교정 단계라고 한다. 교정 단계에서는 근접 해의 변형 부분을 찾는 단계와 변형 부분을 실제로 수정하는 작업이 필요하다. 실제적인 적용 작업에서 다시 예외가 발생하면 이 예외를 처리하기 위해 지식베이스의 룰 지식(rule knowledge)을 이용한 통합 작업이 실시된다.

- 단계 4 : 평가 및 검증

교정 단계를 거쳐 제시된 해와 질의어를 다시 비교 평가하는 단계로 재귀적 호출(recursive call)이 발생한다. 검증 단계에서 제시된 해와 질의어의 평가를 통해 일치성이 확인되면 사례베이스 보완 단계로 이동하고 제시된 해와 질의

어가 맞지 않으면 그 원인을 분석하기 위해 교정 단계로 되돌려진다. 또한 이 과정에서는 검증의 실패원인을 설명하고 학습과정에서 수정되지 않는 예측적 특성(predictive features)이 있는 새로운 사례는 색인생성규칙 단계로 피드백시켜 다시 색인화를 거치는 사이클이 발생하기도 한다.

- 단계 5 : 사례베이스 보완

평가 및 검증 단계에서 검증과 수정을 마친 새로운 사례는 새로운 문제 해결을 위해 사례베이스에 저장된다. 이러한 사례베이스의 보완을 통해 시스템은 자동적으로 지식의 확장이라는 학습효과가 발생하고 반복적 학습기능을 통해 시스템은 지능화된다.

### 3.3.3 유사도 평가

유사도(similarity)란 완전한 일치와 완전한 불일치 사이에 존재하는 연속적 개념이다. 일치하는 것은 공통된 특성을 지니는 반면, 불일치하는 것은 이질적 특성을 갖게 된다. 실세계의 대상들은 대부분 완전히 일치하는 경우보다 유사한 경우가 많다. 이러한 유사한 정도를 평가하는 유사도 측정은 추론의 경우에 많이 활용된다. 특히 사례기반 추론은 유사도 기반 학습이라 불릴 만큼 유사도는 사례기반 추론의 유용성을 평가하는데 중요한 기준이 된다. 특히 큰 사례베이스에서 사례객체들 사이의 유사도 측정은 매우 까다롭고 어려운 문제이다.

유사도를 측정하기 위한 척도로는 거리 개념이 주로 사용된다. 간단하게는 특정수준에서 과거의 사례가 현재의 사례와 부합할 때 유사하다고 말하며, 설명 수준(descriptor level)에서는 각 항목들간의 거리를 계산하여 유사도를 측정하게 된다. 귀납적 검색방법에서의 유사도 측정은 유사한 사례를 군집화하거나 색인화 하여 적절한 형식으로 사례베이스에 저장하고 분류과정에서 주어진 입력과 비슷한 사례의 부류를 추출하고 추출된 사례의 부류를 입력과의 거

리로 환산하는 과정이 필요하다.

유사도 평가함수를 정의하기 위해 먼저 벡터공간 모델에서 모든 색인어가 서로 독립이라고 가정하면 문서  $d_i$ 는  $d_i = (w_{i1}, w_{i2}, \dots, w_{in})$ 와 같은 벡터로 표현된다. 이때  $w_{ik}$ 는 문서  $d_i$ 에서의 색인어  $t_k$ 에 대한 가중치 값이며 만일 문서  $d_i$ 에 색인어가 나타나지 않으면 가중치는 0이 할당된다. 문서벡터가 형성된 이후 질의 검색과정은 벡터 연산에 의해 이루어진다. 즉 그림 3.6과 같이 문서  $d_i$ 와 질의  $q$ 사이의 코사인 값으로 벡터 유사도가 계산될 수 있다. 이것을 코사인 유사도(cosine coefficient similarity)라고 하고 식 (3.2)와 식 (3.3)으로 정의한다.<sup>[41]</sup>

$$\cos(d_i, q) = \left\{ \frac{d_i \cdot q}{|d_i| \times |q|} \right\} \quad (3.2)$$

$$\text{where } \left\{ \frac{d_i \cdot q}{|d_i| \times |q|} \right\} = \frac{\sum_{k=0}^t (w_{ik} \times q)}{\sqrt{\sum_{k=0}^t w_{ik}^2 \times \sum_{k=0}^t q^2}} \quad (3.3)$$

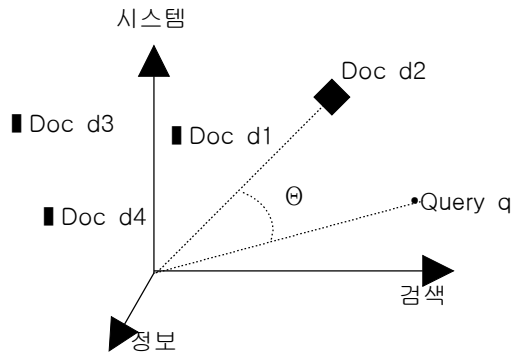


그림 3.6 두 패턴 벡터간 코사인 값을 통한 유사도 평가



유사도 측정에 문서 값은 색인어들의 가중치에 의해 결정되므로 가중치는 검색효율을 결정하는 중요한 요소이다. 가중치 할당기법은 여러 가지 방식이 있지만 웹 정보 검색과 관련해서는 출현빈도, 문서빈도, 정규화의 세 가지 요소가 고려되는 벡터공간 모델이 많이 사용된다.

출현빈도는 문서 내에서 자주 출현하는 색인어에 보다 높은 가중치를 부여한다. 문서빈도는 전체 문서들 중에서 적은 수의 문서에 출현하는 색인어에 보다 높은 가중치를 부여한다. 그리고 정규화는 문서집합에 있는 모든 벡터들의 길이를 같게 만들어 작은 크기의 문서가 가중치 계산에서 불공평하게 취급되지 않도록 한다.

정보검색 벡터모델에서 사용하는 가중치 설정은 단어 빈도수  $TF(t, d)$ 와 문서 빈도수  $DF(t)$ 의 조합으로 나타난다.  $TF(t, d)$ 는 각 문서에 대한 단어  $t$ 가 문서  $d$ 에 나타난 횟수를 의미하고  $DF(t)$ 는 단어  $t$ 가 한번 이상 나온 문서의 수를 나타낸다. 따라서  $k$ 번째 가중치  $w_{ik}$ 는 식 (3.4)와 같이 정의된다. 이때  $\log \frac{|D|}{DF(t_k)}$ 는 역 문서빈도수라하며  $|D|$ 는 문서의 총 개수를 의미한다.

$$w_{ik} = TF(t_k, d) \times \log \frac{|D|}{DF(t_k)} \quad (3.4)$$

가중치  $w_{ik}$ 는 0보다 크거나 같은 값이기 때문에  $\cos(di, q)$  값은 0과 1 사이의 값이 된다. 따라서 연속적 가변 가중치를 사용하는 벡터공간 모델은 문서가 질의와의 관련 여부만을 예측하기보다는 질의와 유사도 값에 따라 순서를 결정할 수 있고 부분적으로 일치되는 문서라도 검색이 가능해 진다.

### 3.4 지능형 검색을 위한 타당성 검토

웹 사용자의 증가는 분산된 웹 데이터의 검색 양을 폭발적으로 증가시키고

있고 이 웹 데이터에는 사용자가 미처 파악하지 못하는 중요한 정보가 포함되어 있으나 이것을 추출할 방법은 거의 없었다. 이것은 웹 데이터베이스를 구성할 때 외부적으로 표출된 표층 지식을 중심으로 스키마를 구성하기 때문에 일정한 형식의 질의를 벗어난 정보 검색에는 대응하지 못하는 데이터베이스 시스템의 한계성을 의미한다. 따라서 숨겨진 심층지식을 추출하기 위한 지능적인 검색 기법이 검토되어야 한다.

표 3.2 지능형 알고리즘 구현을 위한 타당성 검토 항목

방법론	유효성 평가 항목	선택여부
연관규칙 탐사	▪ 숨겨진 심층 지식의 정보 추출 기능	
	▪ 정보 패턴 분석에 따른 재현율 높은 검색 기능	
사례기반 추론	▪ 사용자 선호도에 따른 개인화된 검색 기능	
	▪ 지식획득 및 확충을 위한 사례기반 학습 기능	

본 논문에서 제안한 연관규칙 탐사와 사례기반 추론 기법의 적용 타당성을 표 3.2와 같은 항목을 중심으로 검토해 보면 먼저 연관규칙 탐사는 빈발항목 선정을 통하여 한 항목 그룹과 다른 항목 그룹 사이의 부분적 연관성 여부를 검증할 수 있다. 이제까지 규칙이라 함은 불변의 규칙만을 생각하지만 연관규칙은 사용자가 지정한 지지도와 신뢰도를 만족하는 규칙이면 모두 규칙으로 수용할 수 있다. 이러한 연관규칙으로 숨겨진 심층지식의 정보 패턴을 발견하고 이를 바탕으로 관련성 높은 정보를 제공할 수 있다. 이것은 정보검색의 재현율을 높일 수 있는 보완적 기능을 제공하기 때문에 검색 알고리즘 구현의 기반 기술로 적용할 가치가 높다고 볼 수 있다.

다음으로 사례기반 추론은 과거의 사례를 바탕으로 현재의 문제를 해결하고 현재의 해를 적응시켜 새로운 문제를 해결 할 수 있기 때문에 새로운 상황을

해석하거나 새로운 해에 대한 유효성을 검증하기 적합한 방법론이다. 또한 이 추론 방법은 지식기반 시스템의 문제 해결 방법과 유사하기 때문에 해석, 진단, 유사도 측정 및 학습 분야에 적용이 가능하고, 특히 데이터베이스에 저장된 정보의 추출과 적응을 통해 유사도를 측정하는 검색 에이전트 분야에서는 활용도가 더욱 높다고 볼 수 있다.

사례기반 추론 기법을 검색 알고리즘에 적용하면 유사도 평가라는 문제 해결 방법을 통해 질의어와 완전히 일치하지 않는 검색 결과도 도출할 수 있고 지식확장 같은 사례기반 학습 및 사용자의 선호도에 따른 개인화된 검색 기능도 갖게 된다.

## 제 4 장 질의어 처리 알고리즘의 설계

### 4.1 연관규칙 탐사 알고리즘 설계

#### 4.1.1 빈발항목 생성

연관규칙의 관련성을 결정짓기 위해 가장 많이 사용되는 Apriori 알고리즘에서는 기본적으로 미리 사용자가 정의한 최소지지도 이상의 트랜잭션 지지도를 갖는 빈발항목 집합을 결정하고 이 집합 중에서 빈발항목 요소 상호간에 규칙성을 찾아내어 신뢰도를 생성한다. 따라서 빈발항목의 생성은 요소 상호간의 관련 정도를 결정하는 중요한 평가 기준이 된다. Apriori 알고리즘에서 사용하는 중요한 법칙은 빈도수가 높은 항목 집합의 모든 부분 집합도 빈도수가 높다는 사실이다. 만약 주어진 요소수가  $n$ 개 있을 때 이 항목을 이용해 만들 수 있는 부분집합의 수는  $2^n$ 이다. 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  $k$ 인 항목에서 지지도를 만족하는 집합에 대해서만 요소수가  $k+1$ 인 후보 항목 집합의 지지도를 결정하고 지지도를 미달하는 항목집합은 후보그룹에서 탈락시킴으로써 조합 가능한 부분집합의 수를 줄여나간다.

하지만 본 논문에서 제안하는 연관규칙 추론은 검색엔진의 특성상 주어진 요소의 조합이 많아야 두 개 이상을 넘지 않는다는 제약이 있기 때문에 Apriori 알고리즘과는 다른 방법을 고려해야 한다. 즉  $n$ 개의 집합이 있다면 이 항목을 이용해 만들 수 있는 순서적 의미를 지닌 두 요소 항목 조합은  $n(n-1)$ 개이다. 예를 들면 질의어 요소 수가 3개인 집합에서 순서적 의미를 갖는 두 요소 항목 조합은  $(a, b), (a, c), (b, a), (b, c), (c, a), (c, b)$ 이다. 이때 첫 요소 항목이 빈발항목 생성과 지지도 계산에 기준 요소가

됨으로 첫 번째 항목을 중심으로 그룹화하는 작업을 실시한다. 그룹화를 통해 빈발항목집합을 생성하는 알고리즘을 그룹화 규칙 생성 알고리즘이라 정의하고 그림 4.1에 나타내었다. 그림 4.2는 사용자가 입력한 항목에 빈도수와 항목조합에 따라 지지도를 계산하고 기 정의된 최소 지지도(preset)에 따라 후보 항목 집합이 선정되는 과정을 예시하고 있다. 그림 4.2에서 조인 연산에 의해 생성된 순서적 의미를 갖는 항목 조합 트랜잭션인 (a, b)항목은 규칙 {a} -> {b}를 의미하며 이것은 규칙의 좌항이 규칙의 우항과 직·간접적으로 관련을 갖는다는 것을 의미한다.

```

Procedure Apriori-S(input  $L_k$ , output  $A_k$ ) {
  Ak=null;
  insert into Ck
    select p.item1, q.item1
    from Lk p, Lk q
    where p.item1 <> q.item1;
  forall transaction  $t \in C_k$  do
  {
    t.count ++;
  }
  forall subsets  $s \in t$  do
    forall items first_item  $\in s$  do
    {
       $A_k = \bigcup_k (s.count);$ 
    }
  if  $P(A_k) > supp(min)$  then
  {
    return Ak;
  }
  else
  {
    delete t from Ck;
  }
}

```

그림 4.1 그룹화 규칙 생성 알고리즘

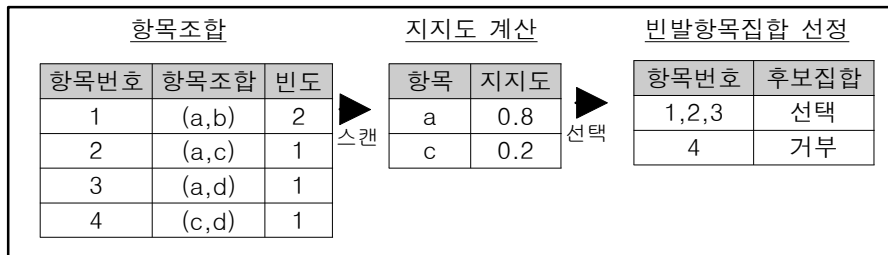


그림 4.2 그룹화를 통한 항목조합 지지도 계산

또한 전체 트랜잭션에 대한 항 {a}에 대한 항 {b}의 관련성을 확률로 나타낼 수 있음을 의미한다. 좌 항을 기준으로 합집합 연산이 수행되면 항 {a}와 관련된 항 {b}, 항 {c}, 항 {d}의 지지도의 합은 전체 트랜잭션에 대한 항 {a}에 대한 지지도를 확률 값으로 나타내며 이 값은 최소 지지도인 임계값에 의해 후보 항목 집합으로 선정되거나 혹은 거부된다.

#### 4.1.2 신뢰도 평가함수

초기 단계의 방대한 항목조합에서 빈발항목집합이 선정되면 두 번째 단계로 항목 상호간의 관련 정도를 결정하는 신뢰도 평가함수를 구현해야 한다. 신뢰도의 평가는 지지도를 만족하는 빈발항목집합 중에서 항목 요소 상호간의 불린 연산이 AND연산인지 혹은 OR연산인지에 따라 다른 연관 가중치가 주어지며 항목사이의 연산 횟수를 통해 신뢰도 평가함수가 결정된다. 이 신뢰도의 결과 값은 기 정의된 임계값인 최소신뢰도에 따라 최종 유효 연관 집합으로 선택되거나 혹은 거부된다.

그림 4.3은 임계값 이상의 지지도를 갖는 후보 집합에서 각 항목 요소 상호간의 관계를 나타낸다. 항목 요소간 관계 연산이 AND인 경우 실선으로, OR인 경우는 일점쇄선으로 표현하였다. 여기서 두 항목 요소간 관계 연산에 따라 다른 가중치를 주었으며 각 후보 집합의 신뢰도  $Rs$ 는 식 (4.1)을 통해 결

정된다.

$$\text{신뢰도}(R_s) = (\alpha \cdot \frac{S_i \rightarrow S_j \text{의 항목수}}{S_i \text{ 항목수}} \cdot A_n) + (\beta \cdot \frac{S_i \rightarrow S_j \text{의 항목수}}{S_j \text{ 항목수}} \cdot O_n) \quad (4.1)$$

식 (4.1)에서  $\alpha$ 는 AND 연산의 가중치,  $\beta$ 는 OR 연산의 가중치를 의미하며,  $A_n$ 은 AND 연산의 횟수 그리고  $O_n$ 은 OR 연산의 횟수를 의미한다. 후보 집합 내 항목간의 신뢰도가 제시된 임계값 이상을 만족하는 경우에 두 항목간에 관련성이 있다고 정의할 수 있다. 표 4.1은 그림 4.3을 식 (4.1)에 따라 항목요소 상호간의 신뢰도를 계산하여 유효 연관 집합의 선택여부를 나타내고 있다. 표 4.1의 평가 예에서는 AND 연산의 가중치는 1로 OR 연산의 가중치는 0.5를 부여했고 최소지지도는 0.3을 부여한 경우이다. 이 표에서 기호 ‘√’는 선택을, 기호 ‘X’는 거부를 의미하고 기호 ‘-’는 해당사항 없음을 의미한다.

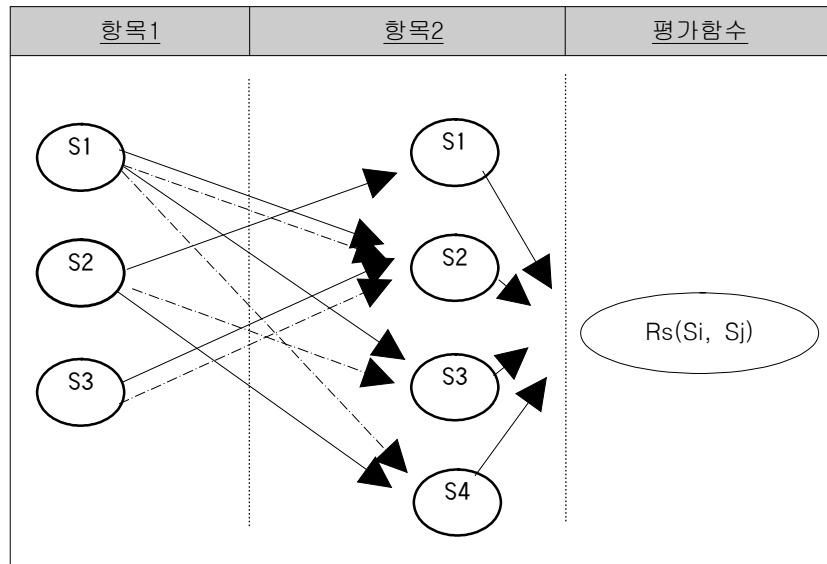


그림 4.3 카티전 프로덕트를 통한 신뢰도 계산

표 4.1 그림 4.3을 이용한 신뢰도 평가 예

요소1 \ 요소2	S1		S2		S3		S4	
	신뢰도	선택	신뢰도	선택	신뢰도	선택	신뢰도	선택
S1	-	-	0.38	√	0.25	×	0.13	×
S2	0.33	√	-	-	0.17	×	0.33	√
S3	0.0	×	0.75	√	-	-	0.0	×

## 4.2 사례기반 추론 알고리즘 설계

### 4.2.1 사례기반 추론 절차

본 논문에서 제안하는 사례기반 추론 알고리즘은 인간이 문제를 이해하는 과정에서 사용하는 추론방법으로 지식이 전문화되고 세분화 될수록 특정한 질의어를 의미론적으로 이해하기 위해 확률적으로 관심 빈도가 높은 분류집합(classification set)에 속하는 색인어가 그 질의어와 관련성이 높다는 것을 전제(premise)한다. 즉 특성에 따라 정보들이 미리 분류되어 있고 주어진 질의어와 관련된 정보가 분류 집합 안에 포함되어 있다면 사례적 통계로 볼 때 방문한 빈도가 높은 분류에 속한 하부 정보들이 그 질의어와 관련성이 높다고 볼 수 있다. 이것을 인지적 확률모델이라 부르며 이 확률모델이 정보의 지역성을 결정하게 된다.

추론 절차는 그림 4.4에서 보인 바와 같이 먼저 새로운 질의어가 주어지면 개인사례베이스의 사례를 이용하여 유사도를 계산하고 카테고리 그룹을 결정하여 질의어와 일치하는 카테고리 그룹의 하부 트랜잭션을 관련 해로 제공한다. 만약 일치하는 색인어가 없으면 일반사례베이스를 검색하여 유사 해를 제공한다. 이 유사 해의 유효성 정도가 주어진 임계값 이상이면 개인사례 테이블에 새로운 사례로 저장한다.



만약 일치하는 유사해가 일반사례베이스에도 존재하지 않는다면 검색 로봇을 트리거(trigger)하여 웹에서 패턴 비교된 일반 하이퍼텍스트 데이터를 사용자에게 제공한다. 사용자가 특정한 하이퍼링크 정보들을 선택하면 선택된 항목을 기준으로 새롭게 유사도를 평가하는 적응 단계를 수행한다. 적응 단계를 거친 유사 해는 다시 사용자로부터 연관성을 평가받아 임계값 이상의 유효성이 검증되면 개인사례베이스와 일반사례베이스에 유사 사례로 저장된다. 일반사례베이스의 유사도 평가는 모든 일반사용자의 질의어 관련 사례 정보를 이용한다.

```

Procedure case_based_algorithm {
  insert into query_String;
  make case_index through similarity;
  select category_group from private_case_table where
  query_String like case_index;
  if forall query_String ∈ case_index then
    retrieved_solution ← subItems of category_group;
  else
  {
    make case_index through similarity;
    select category_group from public_case_table where
    query_String like case_index;
    if forall query_String ∈ case_index then {
      retrieved_solution ← subItems of category_group;
      if retrieved_solution > threshold then
        insert into private_case_table;
        insert into public_case_table;
    }
  }
  else
  {
    pattern data ← web data of Robot_Agent;
    make case_index through similarity with choose items;
  }
}
}

```

그림 4.4 인지적 확률모델 기반의 사례추론 알고리즘

#### 4.2.2 사용자 모델링을 통한 사례기반 학습

사용자 모델링이란 한 시스템이 자신의 목적을 보다 효과적으로 수행할 목적으로 에이전트가 사용자의 독특한 행동을 파악하고 사용자의 행위를 계속적으로 관찰하여 필요한 사용자별 행동패턴을 습득하고 추상화하는 과정이다.

사례기반 학습은 주어진 문제를 해결할 때 과거의 사례를 해로 이용하는 데 대부분의 경우 현재의 문제에 정확하게 일치하는 해는 거의 없기 때문에 과거 사례에 대한 해를 새로운 상황에 적합하도록 변형해야 한다. 이 과정을 교정 혹은 적응이라고 하며 적응을 통해 검증된 새로운 사례의 발견을 사례기반 추론시스템에서는 학습과정이라 한다.

정보검색에서 사용자 모델링을 통한 사례기반 학습은 개인 사례베이스에서 일치하는 사례가 발견되지 않으면 관련성이 낮은 일반 사례베이스에서 사용자에게 유사한 사례 집합을 의도적으로 제시하여 정보검색 에이전트를 학습시키는 것이다. 즉, 사용자가 정보검색 에이전트에게 과거의 사건, 상황에 대한 일반적인 사례를 제시하고 그러한 상황에서는 무엇을 해야 할 것인지를 보이면서 정보검색 에이전트를 훈련시키는 것이다. 정보검색 에이전트는 새로운 사례와 기존사례 사이의 상관관계를 계산하고 사례베이스를 적절히 변화시키면서 새로운 사례를 수용한다.

본 논문에서 제안하는 사용자 모델링을 통한 사례기반 학습은 그림 4.5와 같이 사례표현 단계에서 질의어가 주어지면 질의와 관련된 사용자 속성을 사례베이스에서 가져와 방문 빈도가 높은 카테고리별로 순서화 한다. 작업메모리에서 관련성이 낮은 정보들은 잘려지고 필터링 된다. 사례회수 단계에서는 유사도를 평가하여 유사성 정도에 따라 카테고리 그룹들을 결정하고 카테고리 하부의 정보를 사용자에게 제공한다. 적응 단계에서는 사용자가 일반 사례 집합으로부터 추출된 정보들 가운데 어떤 정보에 관심을 더 보이는가에 따라 유사성의 정도가 강화되기도 하고 약화되기도 하며, 이 정보는 개인 사례베이

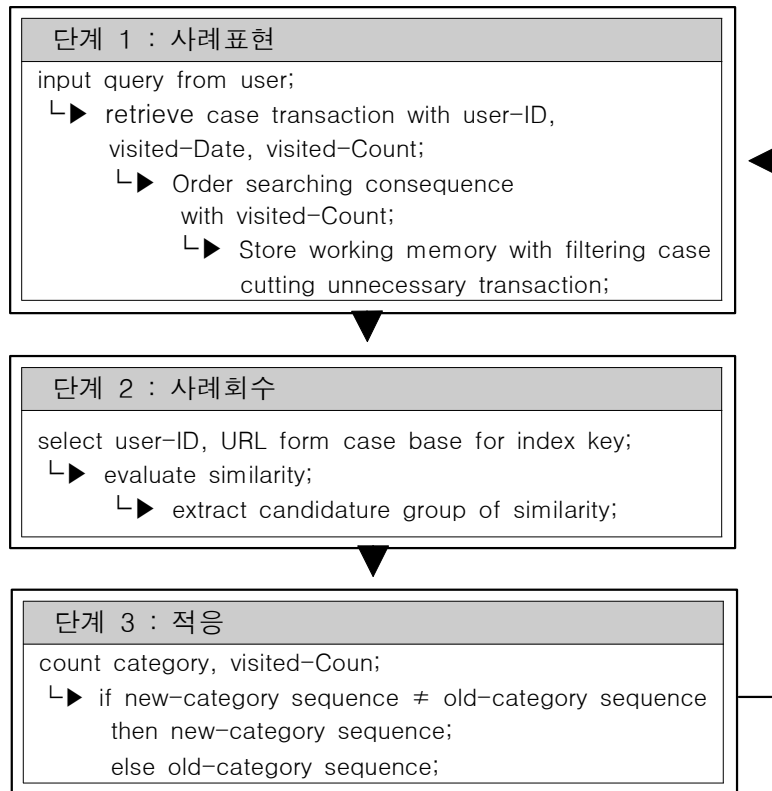


그림 4.5 사용자 모델링을 통한 사례기반 학습 과정

스에 새로운 사례로 저장되어 다시 유사도를 계산하는데 이용된다.

적응단계에서 여과된 정보를 피드백 시키면 정보는 더욱 세분화되고 지식은 계속 확장되어 질 수 있다. 이렇게 사용자 모델링을 통한 사례기반 추론은 사용자의 사례그룹을 계속 강화시키는 학습을 가능하게 한다.

#### 4.2.3 유사도 평가 방법

사례기반 추론에서 저장된 히스토리와 질의간의 완전히 일치하는 사례를 찾지 못할 경우 부분적인 일치를 허용하게 되는데 이 부분적인 일치 즉 유사성을 어떻게 평가하느냐에 따라서 시스템의 성능이 좌우될 수 있다. 사례를 평

가하는 전통적인 유사도 평가 방법으로 불린모델, 벡터공간모델, 확률모델이 있다. 불린모델은 집합론과 불리안 대수학에 기반하는 간단한 집합이론에 근거하고 있으며 문헌과 질의가 색인어의 집합으로 표현되고 집합과 집합 연산자로 구성되어 있다. 이 방법은 직관적이고 이해하기 쉽고 사용자 요구인 질의의 의미가 명확하게 전달 될 수 있지만 색인어 가중치( $w_{ik}$ )를  $\{0, 1\}$  값만 가져 연관된 문서인지 아니지만 예측하고 순위화는 할 수 없는 문제점을 지니고 있다.

벡터공간모델은 질의와 문서가 벡터로 표시되고  $n$ 차원 공간상에서의 상대거리를 기준으로 검색을 한다. 이진 가중치의 제한성을 극복하기 위해 가중치를 실수(float-point)로 부여하여 부분 정합과 질의에 유사 정도에 따라 검색된 문헌을 순위화 할 수 있다. 또한 클러스터 내 유사도와 클러스터간 유사도 조정을 위해 용어 가중치를 사용할 경우 검색 성능을 향상시킬 수 있다. 그러나 색인 용어들간의 연관성을 고려하지 않아 의미론적인 검색이 어렵다.

벡터공간 모델 중 한 기법인 최근접 이웃분류법은 새로운 문체의 특성과 사례베이스에 있는 각 사례들과 대응되는 특성을 하나씩 비교하는 매우 간단한 방법이지만 사례베이스의 크기가 증가함에 따라 비용이 급속하게 증가하는 소모적 평가 방법이다.

따라서 본 논문에서는 유사도 평가함수를 위해 통계모델의 일종인 유사 범주화 알고리즘을 제안한다. 이 알고리즘은 사전(a priori) 분류체계에 기초하여 대상 문서를 분류하고 문서나 질의가 주어지면 가장 적합한 범주에 새로운 사례로 할당함으로써 문서들의 집단을 형성하는 기법이다. 즉 이 기법은 사용자 질의어 방문 패턴을 분석하여 빈도가 많은 카테고리별 순위를 결정하고 이것을 사례로 제공하여 같은 이름의 질의어라도 사용자가 의미하는 카테고리 정보의 연관성을 높여 의미론적 접근이 가능하도록 하는 것이 기본 전략이다. 또한 개인 사례베이스에 일치하는 사례 해가 존재하지 않으면 유사사례 추출

을 위해 일반 사례베이스에서 질의어와 일치하는 유사 사례 카테고리를 결정하여 유사 해로 제시하고 사용자로부터 적응과정을 거치도록 하여 새로운 사례로 개인 사례베이스에 저장하도록 한다.

유사 범주화 알고리즘의 평가함수는 질의어  $q$ 에 대해 개인 사례베이스  $pDB$ 의 관련 문서의 수인  $|pDB(q)|$ 와 카테고리  $k$ 에 대한 방문 빈도수  $|C_k(q)|$ 의 확률을  $P(C_k/pDB)$ 로 정의한다. 이때  $|pDB(q)| = \sum_{i=0}^n (q \in T_i, C_i, D_i)$ 로 표현하며 이는 질의어  $q$ 에 대해 개인 사례베이스 내의 제목  $T_i$ , 카테고리  $C_i$ , 서술항  $D_i$ 을 갖는 트랜잭션들과 패턴 매칭 작업을 반복적으로 실시할 때 질의어와 일치하는 트랜잭션들의 수를 의미한다.  $P(C_k/pDB)$ 의 값이 크다는 것은 주어진 예제 질의어에 대해서 사례베이스가 관련 문서를 많이 반환하는 경우로 그 질의어에 대한 사례집합의 유사도가 증가하고,  $P(C_k/pDB)$ 의 값이 작을 경우는 그 질의어에 대한 사례집합의 유사도는 감소한다는 것을 뜻한다. 카테고리 집합을  $C_i$ 라하고 질의어  $q$ 가  $q \subseteq C_i$ 을 만족하는 개인 및 전체 문서 데이터베이스와의 유사도  $SM(q, mDB/m=p, g)$ 을 계산하는 평가함수를 다음 식과 같이 정의한다.

$$\text{유사도}(SM) = P(C_k/pDB) + \omega \times P(C_k/gDB) \quad (4.2)$$

where  $0 \leq \omega \leq 1$

식 (4.2)에서  $P(C_k/gDB)$ 는 질의어  $q$ 에 대해 일반 사례베이스  $gDB$ 가 반환하는 관련 문서의 수인  $|gDB(q)|$  대해 카테고리  $k$ 에 대한 방문 빈도수  $|C_k(q)|$ 의 확률을 의미한다. 개인 사례베이스에서 유사도를 구하지 못할 경우 적응을 위해 일반사례베이스로부터 유사 사례의 추출에 이용되며 가중치  $\omega$ 의 값에 따라 실효치가 달라질 수 있다.

본 논문에서 제안하는 유사도 평가함수는 정보 검색 문제를 확률적 틀로 해

석하고 카테고리 내 유사도는 방문 빈도수(hit frequency)를 통해 표현하고 카테고리 간 유사도는 사례베이스가 반환하는 문서의 수를 통해 나타낸다. 이러한 확률모델을 기반으로 한 유사도 평가함수는 사례집합을 연관확률에 따라 카테고리별로 순서화할 수 있고 선택된 카테고리 그룹의 모든 하부 사례 트랜잭션들은 연관성 높은 정보로 제공될 수 있다. 따라서 사례베이스에 대한 연관도를 조정하는 학습과정이 계속적으로 이루어진다면 사례기반 추천엔진의 지능적 검색 기능은 증가 될 수 있다.

## 제 5 장 지능형 에이전트 시스템의 구현

### 5.1 지능형 에이전트 시스템의 구성

본 논문에서 구현한 지능형 에이전트 시스템(AI-SEA)은 해양정보 전문검색 엔진으로 개발언어는 JSP를 사용하였고 색인 데이터베이스 및 관련 데이터베이스는 오라클 8.1을 사용하여 JDBC를 통해 연동하였다.

그림 5.1은 AI-SEA의 질의 결과 화면이며, 사용자가 질의어 '운반'을 텍스트 필드에 입력하면 추론엔진이 질의어와 부합된 키워드를 이용하여 연관규칙 탐사와 사례기반 추론 기법으로 색인데이터베이스에서 관련 정보를 찾아 HTML 문서 형태로 결과를 보여주는 화면이다.

시스템의 검색 방법은 지능적 검색 정도에 따라 사례기반 검색, AI-SEA 일반검색, 전문웹 검색, 웹페이지 검색으로 나누어진다. 사례기반 검색은 사례기반 추론에 따라 검색하여 전문성과 관련성이 높은 정보를 제공하며 AI-SEA 일반검색은 해양과 관련된 전문적 정보만을 제공한다. 전문 웹 검색은 한미르, 야후, 네이버와 같은 범용검색엔진에서 카테고리 정보를 기준으로 검색하여 사용자에게 정보를 제공하고 웹페이지 검색은 상용검색엔진에서 검색어와 관련된 웹 문서를 검색하여 일치하는 모든 정보를 제공한다. 이 검색 방법은 관련 문서의 양이 매우 많지만 관련성이 없거나 중복되어 있는 문서도 함께 제공될 수 있다. 연관추론은 연관규칙탐사 기법에 따라 질의어의 연관성이 검증된 경우에만 자동적으로 연관어를 함께 제공하는 검색 방법이다.

시스템의 개념적 구조는 그림 5.2와 같이 검색부, 로봇부, 무선부, 추론부로 구성되어 있다. 로봇부는 로봇모듈과 인덱서모듈로 구성되며 추론부는 연관규칙 추론부와 사례기반 추론부로 구성된다. 또한 무선부에 대한 구현기술 및 접속절차에 대해서는 5.5절에서 기술하고자 한다.



그림 5.1 AI-SEA 시스템의 질의 결과 화면

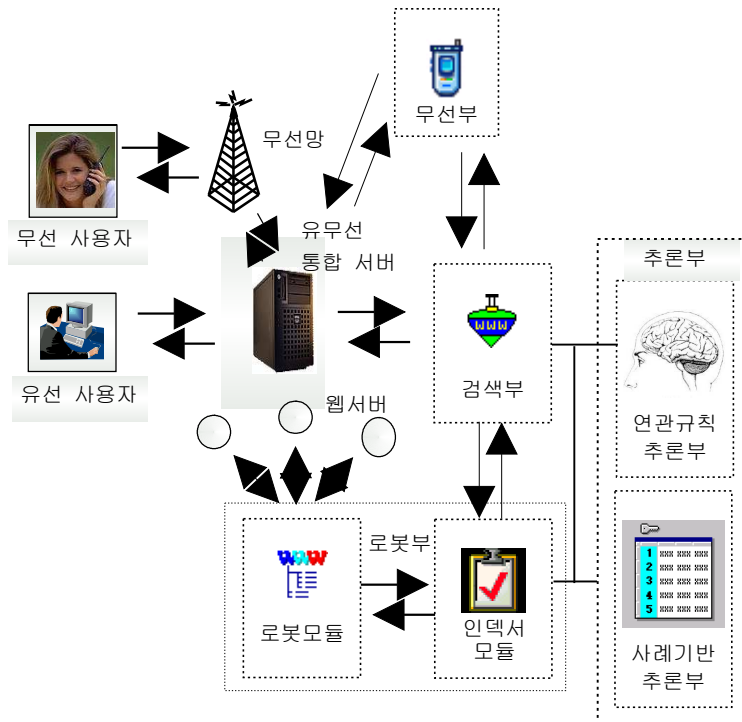


그림 5.2 AI-SEA 시스템의 개념적 구조



## 5.2 로봇부 및 검색부

로봇부와 검색부는 검색 에이전트 시스템의 기본 구성으로 로봇부는 주기적으로 혹은 검색부의 요청이 있을 경우 정보를 추출하는 로봇 모듈과 색인데이터베이스를 구성하는 인덱서 모듈로 구성되어 있다. 검색부는 사용자의 질의어를 검색하여 표현하는 프리젠테이션 계층과 로봇부와 연동된 트리거 계층으로 구성된다. 프리젠테이션 계층은 사용자의 질의어와 색인데이터베이스의 색인어를 비교하여 일치하면 해당 정보를 HTML 문서 형태로 표현하여 사용자에게 보여준다. 트리거 계층은 로봇부와 상호구동적으로 동작하는 계층으로 질의어와 일치하는 색인어가 색인데이터베이스에 존재하지 않을 경우 로봇부에게 웹에서 관련된 내용을 검색하도록 요구한다.

### 5.2.1 로봇부의 구조

검색 로봇의 구현은 여러 가지가 있지만 보통 W3C(world wide web consortium)의 레퍼런스 라이브러리를 참조하여 C, Perl, JAVA와 같은 프로그래밍 언어를 사용하여 제작하고 있다. W3C는 웹의 여러 가지 정책을 세우고 표준을 주도하는 곳이라 CERN서버와 API를 제공하고 있다. 검색 로봇과 관련된 라이브러리는 다음과 같은 두 가지 버전이 사용되고 있다.

- libwww 2.17 : C로 작성되어 있으며 싱글쓰레드 구조로 라이코스가 처음 검색 로봇을 만들 때 이 라이브러리를 사용했다.
- libwww 4.0 : 최신 HTTP 라이브러리로 C로 작성되었으며 이벤트 기반의 멀티 쓰레드를 제공한다.

검색 로봇은 웹의 리소스 발견 및 웹서버의 유지보수와 같은 영역에 폭넓게 사용될 수 있는 자동화된 도구이지만 다음과 같은 몇 가지 문제점을 지니고

있다. 첫째, 검색 로봇 자체가 악용될 소지가 있기 때문에 접근에 대한 표준만 제안되어 있고 libwww를 포함하여 현재 어디에도 완벽한 검색 로봇 프로그램을 공개하고 있지 않아 멀티 에이전트의 통합을 어렵게한다. 둘째, 웹 서버에 'robots.txt'와 같은 로봇 배제 표준안이 제공되지 않는 경우에는 특정 URL을 반복적으로 접근하여 네트워크 기능을 지연시키거나 마비시킬 위험성이 있다.

이러한 문제점을 보완하기 위해 메타 검색엔진에 의한 검색 방법이 제시되었다. 메타 검색엔진은 여러 검색엔진을 동시에 구동함으로써 자체 검색엔진이 부담할 네트워크 병목현상을 분산시켜 효율적인 검색을 수행하고 보다 광범위한 검색 결과를 제공하지만 여러 검색엔진을 구동시키기 위해 많은 자원을 필요로 하고 검색요구가 동시 다발적으로 발생할 경우 오히려 병목 현상을 가중시킬 위험성이 있다.

따라서 본 논문에서는 그림 5.3과 같이 로컬 검색 로봇을 이용한 검색 방법과 메타 검색엔진에 의한 검색 방법의 이점을 상호 보완한 통합 검색엔진(integrated search engine)에 의한 검색 방법을 제안한다. 통합 검색엔진에 의한 검색은 로컬 로봇 배제 파일과 유사한 데드링크(dead link) 테이블을 웹 서버가 아닌 검색 로봇에 자체적으로 보유하여 문제가 있는 사이트의 방문을 제한하여 불필요한 방문으로 인한 병목현상과 데드 링크된 사이트를 배제시켜 검색 결과의 신뢰성을 높인다.

항해 전략은 로우데이터테이블에서 획득한 URL이 저장된 URL테이블에 존재하는 사이트인지 아닌지에 따라 다른 탐색 방법이 적용된다. 먼저 URL테이블에 존재하지 않는 사이트에 대해서는 검색모듈의 요청에 따라 메타 검색엔진과 같이 여러 범용검색엔진에 질의를 요구한다. 질의 요구 순서는 각 범용 검색엔진에 동시다발적으로 정보를 요구를 하지 않고 순차적으로 검색을 의뢰하여 검색된 문서 수가 많은 검색엔진의 정보를 지정된 형식에 맞추어 출력함

으로써 추가적인 처리 지연을 막도록 하였다.

또한 URL테이블에 존재하는 사이트는 검색 로봇이 주기적으로 웹서버를 방문하여 관련 정보를 수집하거나 데드링크 여부를 확인하여 색인데이터베이스를 재구성한다. 로봇부의 동작과정은 그림 5.4와 같이 로봇 모듈이 URL테이블을 참조하여 웹서버로부터 원시자료를 수집하고 이 자료를 인덱서모듈에 전달하여 색인데이터베이스를 구현한다. 만약 URL테이블에 의해 구현하지 못한 색인 정보는 검색부에 의해 트리거되어 웹으로부터 관련 정보를 직접 가져와서 사용자 인터페이스에 제공된다.

인덱서 모듈이 웹 문서의 패턴을 분석하여 URL, 제목, 내용 등의 정보를 추출하고 색인 데이터베이스를 구축하는 과정은 다음과 같이 네 단계로 처리되며 그림 5.5에 자바코드로 나타내었다.

- 단계 1 : URL 데이터베이스에 등록된 도메인 이름을 이용하여 URL 클래스 객체를 생성한다.
- 단계 2 : URL 클래스 객체를 네트워크상의 논리적인 입력 스트림으로 사용하기 위해 버퍼링이 가능한 `InputStreamReader` 객체를 생성한다.
- 단계 3 : 입력 스트림이 `Null`이 될 때까지 반복적으로 문서를 분석하여 문자열 인스턴스 값의 위치정보를 검출한다.
- 단계 4 : 단계 3에서 확인된 위치 정보 값을 이용하여 문자열 인스턴스를 부분 문자열로 잘라내어 `url`, `title`과 같은 저장 정보로 변환하여 색인데이터베이스에 저장한다.

이러한 처리 과정을 통해 질의어와 카테고리 별로 분류된 데이터를 빠르게 데이터베이스화 할 수 있다.

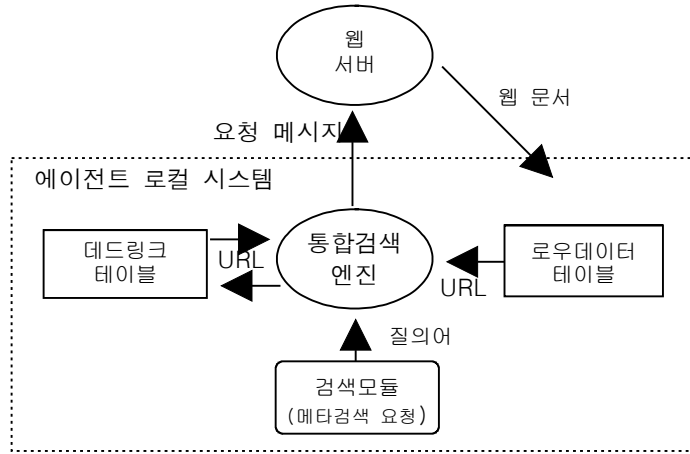


그림 5.3 로봇부의 개념적 구조

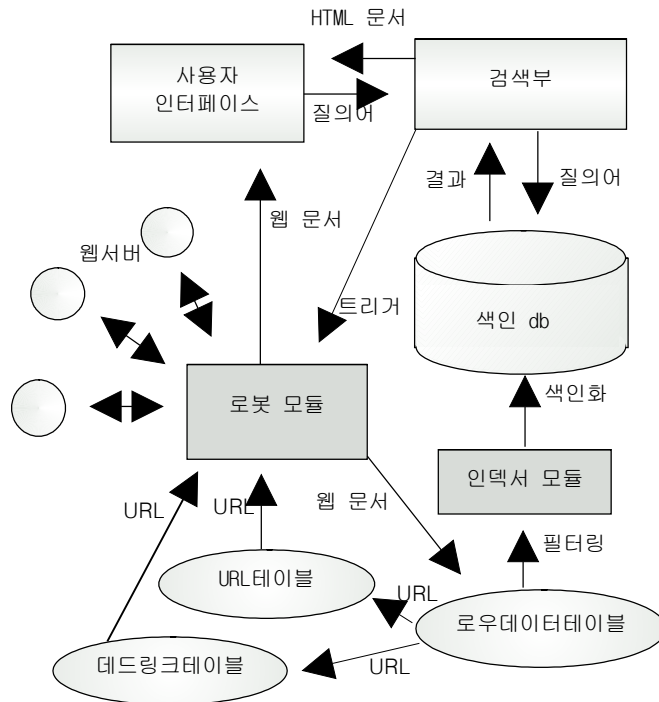


그림 5.4 로봇부의 세부 동작 과정

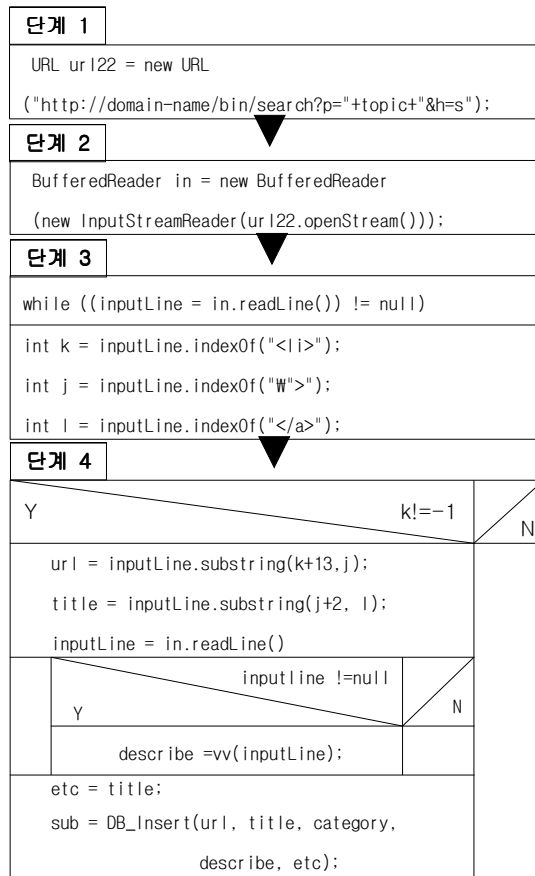


그림 5.5 로봇부의 색인데이터베이스 구현

### 5.2.2 검색부의 구조

검색부는 사용자와 색인데이터베이스 사이에서 상호 동작을 통해 정보의 입출력을 지원하는 인터페이스 모듈이다. 인터페이스 모듈은 색인데이터베이스의 IP 어드레스와 포트번호를 객체화하고 사용자가 질의한 검색어를 SQL언어로 캡슐화하여 원격에 분산된 데이터베이스 서버와 연동한다.

본 논문에서는 색인데이터베이스 드라이버와 내부 연결에 이용되는 JDBC 드라이버로 Oracle 회사에서 제공하는 Thin Driver를 이용한다. 그러나 색인데이터베이스와 사용자간의 실제적인 연결은 “java.sql” 패키지와 “java.net”

패키지에 포함된 여러 클래스의 API 메소드를 통해서 이루어진다.

그림 5.6은 검색부가 색인 데이터베이스 및 로봇부와 연동되는 과정을 보여주고 있다. 자바기반의 검색부는 데이터베이스와 연동하는 과정에서 기본적으로 JDBC를 사용한다. JDBC란 자바를 이용한 데이터베이스 접속과 SQL문장의 실행 그리고 그 실행결과로부터 얻어진 데이터의 처리 방법과 절차에 관한 프로토콜이다. 이 프로토콜은 하부 데이터베이스와 애플리케이션 사이의 연동을 위해 표준 API를 이용하기 때문에 특정 데이터베이스와 독립된 프로그램의 구현을 가능하게 한다.

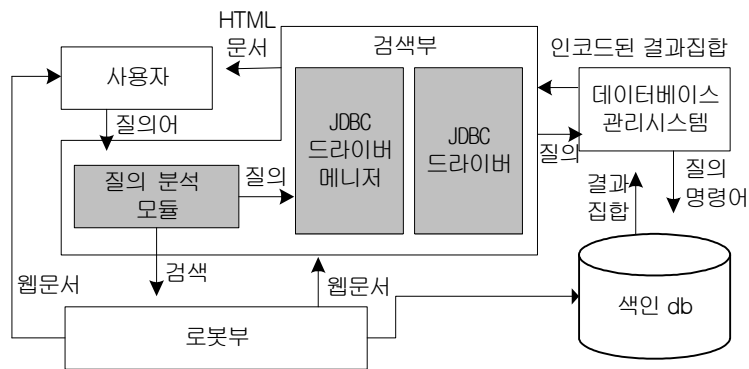


그림 5.6 검색부의 색인데이터베이스 연동 모델

즉 JDBC와 연동이 가능한 드라이버와 모듈을 동적으로 적재해 하부 드라이버의 종류에 상관없이 데이터베이스 소스에 액세스할 수 있는데 이 기술은 클래스로 구현된 드라이버 매니저를 통해 직접 제어할 수 있다. 질의 분석모듈(query analyzer module)은 통합형 SQL 데이터 질의어를 처리하는 프레임워크로서 질의 및 처리결과를 분석하여 색인 데이터베이스 및 로봇부로 제어를 스위칭시키는 역할을 담당하는 제어모듈이다. JDBC 드라이버 매니저(JDBC driver manager)는 데이터베이스별 드라이버를 등록하고 관리하는 모듈로 사용방법은 Class.forName() 메소드를 이용하는 방법, driver 객체를 생성하는

방법, registerDriver() 메소드를 이용하는 방법이 있지만 일반적으로 사용하기 쉬운 Class.forName() 메소드를 주로 사용하고 있다. 본 논문에서도 이 방법을 통해 객체 모듈을 생성하고 등록하였다.

JDBC 드라이버 통하여 색인데이터베이스에 검색어를 질의하여 관련된 레코드를 검출하고 HTML 문서 형식으로 결과를 출력하는 구현 과정은 다음과 같이 세 단계로 이루어지며 그림 5.7에 자바 코드로 나타내었다.

- 단계 1 : DriverManager 클래스에 해당 데이터베이스 driver를 등록하고 Class.forName() 메소드를 사용하여 연동에 필요한 해당 드라이버를 로드하고 getConnection() 메소드로 DBMS에 연결될 세션인 Connection 객체를 획득하여 원격의 데이터베이스와 연결한다.
- 단계 2 : Statement 클래스에서는 쿼리문의 실행을 위해서 ResultSet 클래스의 executeQuery() 메소드를 이용하여 SQL 명령문을 실행한다.
- 단계 3 : ResultSet 타입의 객체를 반환 받아 파싱(parsing)하고 레코드 셋(record set)에 저장된 정보를 HTML 코드 형태로 웹 브라우저에 출력한다.

검색부가 수행하는 기능의 대부분은 색인데이터베이스와 연동하여 사용자의 질의어를 처리하고 표현하는 과정인데, AI-SEA 시스템의 검색부가 사용하는 색인데이터베이스 및 주요 관련 데이터베이스의 릴레이션 구조는 그림 5.8과 같다.

릴레이션 1은 방문 URL 테이블로써 로봇부가 방문할 사이트의 URL, 질의

어, 카테고리 정보 등을 포함하고 있다. 릴레이션 2는 해양정보 테이블로써 웹 문서로부터 추출된 해양 관련 사이트의 URL과 카테고리 정보 및 방문 횟수 정보 등을 포함하고 있다. 릴레이션 3은 불용어 테이블로써 해양정보 테이블의 무결성 보장을 위해 해양관련 전문용어들과 사이트 설명 정보 등을 포함하고 있다. 릴레이션 4는 연관규칙 테이블로써 연관규칙 탐사를 위한 키워드 및 사용자 ID 등을 포함하고 있다. 릴레이션 5는 개인 사례테이블로써 개인별 방문 히스토리 정보인 URL, 질의어, 카테고리, 방문횟수, 타이틀 및 방문일자 등을 포함하고 있다.

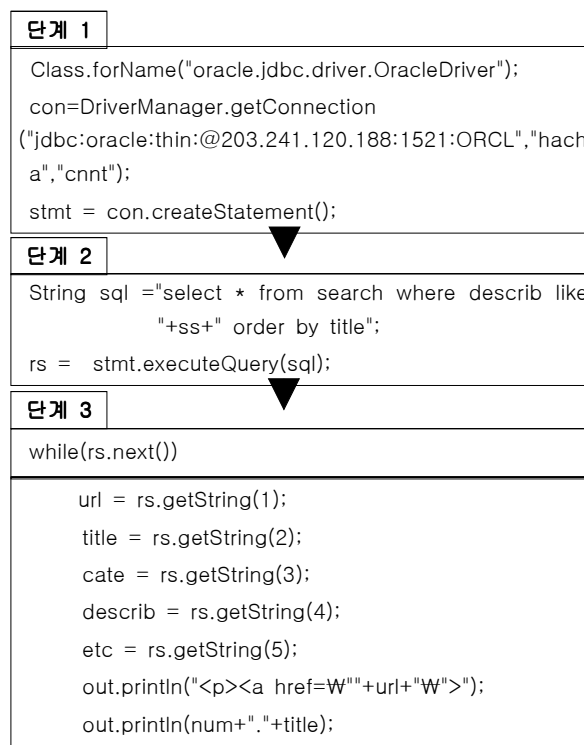


그림 5.7 검색부의 색인데이터베이스 연동 코드



릴레이션 1 : 방문 URL 테이블
<ul style="list-style-type: none"> <li>• 로봣부가 방문할 URL 정보 포함</li> <li>• 테이블명 및 스킴구조 : LINK = (URL, KWORD, CATE, DESCRIB)</li> <li>• URL := IP주소, KWORD :=질의어, CATE := 카테고리 정보, DESCRIB := 사이트 설명부</li> </ul>

릴레이션 2 : 해양정보 테이블
<ul style="list-style-type: none"> <li>• 추출된 해양 관련 사이트의 URL과 카테고리 정보 및 방문 횟수 포함</li> <li>• 테이블명 및 스킴구조 : SEARCH = (URL, TITLE, CATE, DESCRIB, CNT)</li> <li>• TITLE := 사이트명, CNT := 방문횟수</li> </ul>

릴레이션 3 : 불용어 테이블
<ul style="list-style-type: none"> <li>• 해양정보 테이블의 무결성 보장을 위한 해양관련 전문용어 포함</li> <li>• 테이블명 및 스킴구조 : SEA_DIC = (HAN_WORD, DESCRIB)</li> <li>• HAN_WORD := 해양전문용어</li> </ul>

릴레이션 4 : 연관규칙 테이블
<ul style="list-style-type: none"> <li>• 연관규칙탐사를 위한 키워드 포함</li> <li>• 테이블명 및 스킴구조 : WORD_PRIORITY = (WORD, CNT, USER_ID)</li> <li>• WORD := 연관규칙 키워드, USER_ID := 사용자 ID</li> </ul>

릴레이션 5 : 개인 사례테이블
<ul style="list-style-type: none"> <li>• 개인별 방문 히스토리 정보 포함</li> <li>• 테이블명 및 스킴구조 : T_CASE_ADDRESS = (URL, KWORD, TITLE, CATE, CNT, CREAT_DATE)</li> <li>• CREAT_DATE := 최종방문일자</li> </ul>

그림 5.8 색인 데이터베이스 및 관련 데이터베이스의 릴레이션 구조

### 5.3 연관규칙 추론부의 구조

연관규칙 추론부는 검색 트랜잭션에 포함된 아이템들의 속성에 따라 연관규칙을 찾아내는 모듈로 에이전트 시스템에게 연관규칙과 관련된 지능적 검색 기능을 제공한다. 연관규칙 추론부의 주요 구성 요소는 그림 5.9에 표시한 것과 같이 연관규칙 탐사추론 엔진과 연관 지식베이스로 이루어져 있다.

연관 지식베이스는 이전의 검색에서 'AND' 혹은 'OR'의 논리연산자와 함께 주어진 두 개의 질의어에 대한 히스토리 정보를 기록하고 있고 연관추론엔진은 하나의 질의어가 주어졌을 때 기록된 연관어 정보들을 이용하여 최소 지지도와 최소 신뢰도를 만족하는 규칙을 생성하는 부분이다. 사용자 인터페이스를 통해 입력된 두 질의어와 관계연산자는 (Frame\_ID, Operation, Keyword1, Keyword2)와 같은 프레임 형식으로 캡슐화 되어 연관 지식베이스에 표 5.1과 같이 축적된다.

표 5.1 트랜잭션 테이블

ID	키워드
R1	(AND, ELECTRIC PANELS, STARTER BOX)
R2	(AND DOCKING, SHIPPING)
R3	(AND, GOVERNOR, STOPPER)
R4	(OR, MARINE BOILER, TURBINE CONDENSER)
R5	(AND, AIR VENT HEAD, AIR RESERVOIR)
R6	(OR, TURBINE CONDENSER, WATER GENERATOR)

연관규칙 추론부에 최소 지지도와 최소 신뢰도가 그림 5.10과 같이 설정되면 다음과 같이 연관 규칙이 결정된다. 먼저 사용자가 미리 정의한 최소 지지도를 만족하는 데이터 항목 집합을 탐사하기 위해 4장에서 기술한 지지도 평

가함수에 의해 각 데이터 항목에 대하여 지지도가 계산되고 최소지지도를 만족하는 빈발항목집합이 결정된다.

다음으로 추출된 빈발항목 집합들 중에서 4장에서 기술한 최소 신뢰도 평가함수를 만족하는 규칙들을 탐사하여 후보 항목집합이 표 5.2과 같이 최종 대상으로 결정된다.

표 5.2 발견된 연관규칙

발견 규칙	확신도
[MARINE BOILER]→[TURBINE CONDENSER]	0.82
[AIR VENT HEAD]→[AIR RESERVOIR]	0.64
[TURBINE CONDENSER]→[WATER GENERATOR]	0.50
[GOVERNOR]→[STOPPER]	0.45

빈발항목집합과 신뢰도 있는 연관어를 생성하는 알고리즘은 그림 5.11에서 자바코드로 나타낸 것처럼 다음과 같이 세 단계로 구성된다.

- 단계 1 : 검색 창에서 사용자에게 의해 주어진 질의어를 Ck 테이블의 기본 키와 비교하여 일치하는 레코드들을 객체배열에 저장하고 지지도 평가함수를 이용하여 각 레코드 항목의 지지도를 계산한다. 각 항목의 지지도는 임계값으로 주어진 최소 지지도와 비교하여 빈발항목 집합을 결정하고 “cand\_sup” 객체 배열에 그 결과를 저장한다.
- 단계 2 : 신뢰성 있는 연관규칙을 생성하기 위해 1단계에서 추출된 빈발항목집합을 대상으로 신뢰도 평가함수로 각 항목의 신뢰도를 계산하여 “cand\_rec” 객체배열에 그 결과를 저장한다.

- 단계 3 : 객체 배열에 저장된 값인 후보항목 집합의 각 값을 신뢰도 평가함수에 의해 계산하여 임계값으로 주어진 최소신뢰도와 비교하여 최소신뢰도 이상의 값을 갖는 후보항목 집합은 cand\_rec[i] 객체 배열에 저장하여 연관성 정도를 결정하는 자료로 활용한다.

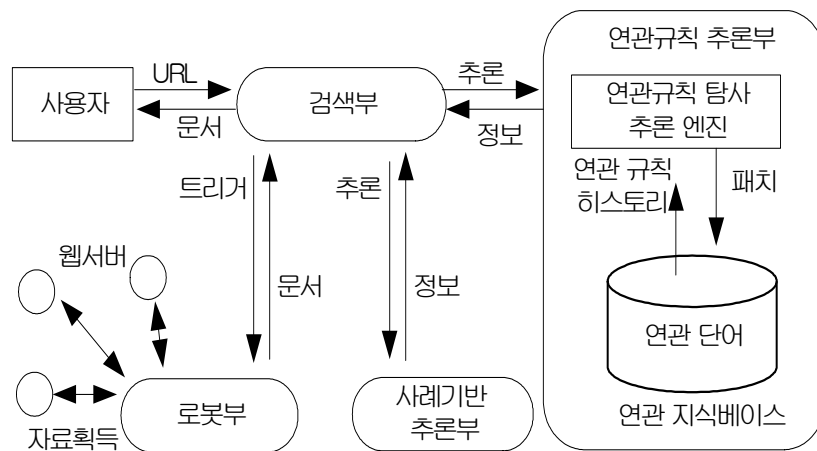


그림 5.9 연관규칙 추천부의 구조

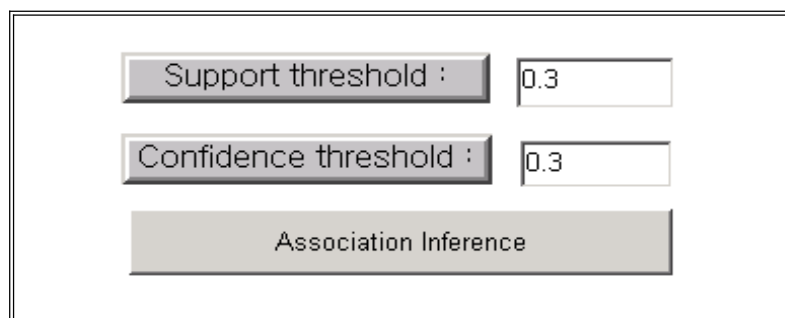


그림 5.10 최소 지지도 및 최소 신뢰도 임계값 지정

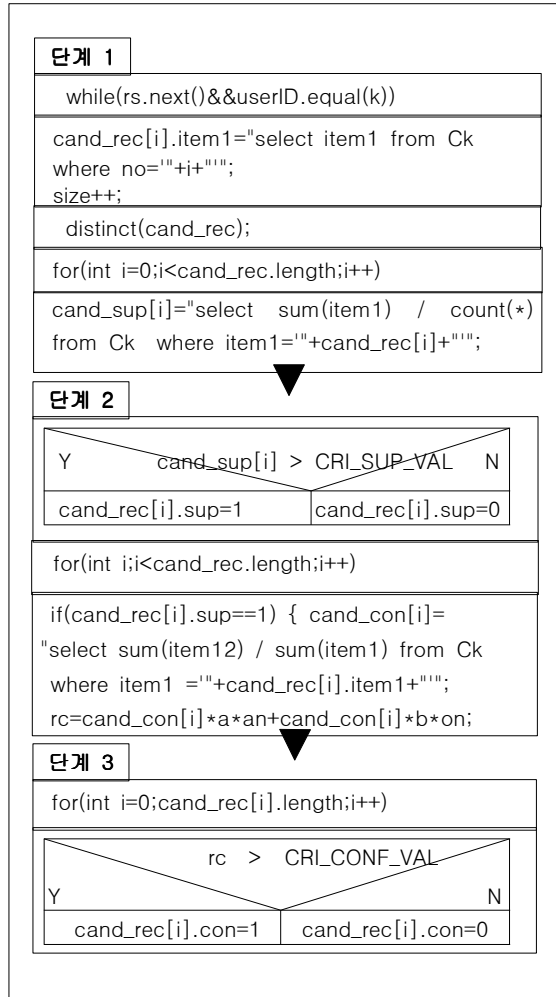


그림 5.11 빈발항목집합 및 연관규칙 생성 코드

그림 5.12는 ‘해양’이라는 질어어가 입력된 경우 연관규칙 탐사 추론을 수행하여 연관성이 45.4% 정도 있는 ‘유람선’이라는 항목을 연관어로 함께 보여주고 있는 화면이다.



그림 5.12 연관규칙 탐사를 통해 연관어가 추출된 검색 결과

#### 5.4 사례기반 추론부의 구조

사례기반 추론부는 그림 5.13에 나타낸 것과 같이 사례지식베이스(case knowledge base)와 유사도측정엔진(similarity measure engine)으로 구성되어 있다. 사례베이스는 이전에 사용자가 검색한 질의어에 따라 방문한 기록들을 저장하는 부분으로 개인 사례베이스, 일반 사례베이스, 등록문서 데이터베이스로 구성되어 있다. 개인 사례베이스는 개인화된 정보 검색을 지원하기 위해 개인방문 패턴을 질의어 별로 기록하고 있는 테이블이며 일반 사례베이스는 등록하지 않은 일반인의 방문 기록을 저장하고 있는 테이블이다. 등록문서 데이터베이스는 사용자의 모든 방문기록을 일정기간 임시적으로 저장하는 저장 테이블이다. 이 테이블은 해양관련 전문 콘텐츠와 관련이 없는 방문기록들을 사례베이스에 저장하기 전에 필터링 하는 전처리 과정을 수행하기 위해 이용된다.

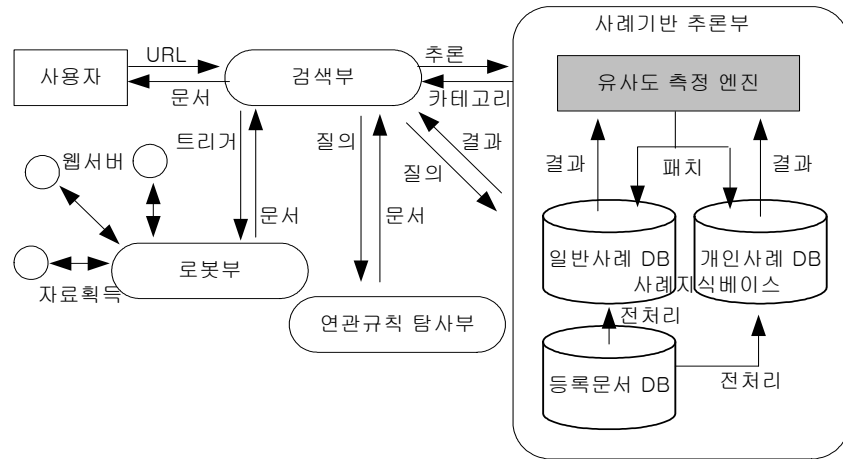


그림 5.13 사례기반 추천부의 구조

유사도 측정 엔진은 질의어가 소속될 카테고리 그룹을 순서화 하는 유사 군집화 프로그램으로 구성되어 있다. 유사도 평가함수에 의한 카테고리 결정은 개인별 맞춤 검색의 이론적 근거가 되기 때문에 개인 사례베이스와 일반 사례 베이스를 이용하여 개인별 카테고리 그룹을 결정하고 이 그룹을 개인별 사례 추론을 위한 기준으로 사용한다.

예를 들어 그림 5.14의 사례별 카테고리 그룹에서처럼 같은 이름의 학과나 연구실이라 하더라도 질의어가 ‘공학및조선’에 있느냐 ‘자연과학’에 있느냐에 따라 다른 내용이 저장되어 있기 때문에 패턴 비교에 의한 단순 비교는 관련성 없는 결과를 추출할 수 있다.

유사도 평가함수에 등록된 개인이나 비등록된 일반의 방문패턴 기록을 함께 고려하는 이유는 개인의 방문 히스토리 기록이 없거나 정보량이 적을 경우에 주어진 질의어에 대한 일반인의 방문 패턴 정보를 함께 이용하기 위한 것인데 개인화된 정보 추출의 정확도를 높이기 위해 일반인의 방문패턴 기록에 대한 가중치는 상대적으로 낮추었다.



그림 5.14 사례별 카테고리 그룹 계층도

유사도 측정엔진은 다시 사례검색기와 사례추론기로 나뉘어진다. 사례검색기는 어떤 문제가 주어지면 사례베이스에서 가장 적절한 사례를 식별하고 유사도를 평가하기 위해 사례추론기를 호출한다. 사례추론기는 사례베이스에 축적된 사례를 조사하여 유사도를 계산하고 필요한 경우 적응과정을 거쳐서 새로운 문제의 해결을 시도한다. 응용분야의 특성이나 사례베이스의 내용에 따



라서는 사례추론기가 필요치 않을 수도 있다. 즉 거의 정확한 사례들을 찾는 경우 이를 사용자에게 그대로 제시하고 사용자는 제시된 몇몇 정보들 중에서 적합한 정보를 선택 할 수 있다. 그러나 대부분의 경우 사례는 질의어와 부분적으로 일치됨으로 사례추론기는 일반적으로 호출된다.

사례베이스의 생성은 다음과 같이 세 단계로 나누어 구현 되며 그림 5.15는 유사도 평가함수의 추론 과정을 자바코드로 나타낸 것이다.

- 단계 1 : 웹 문서로부터 사례지식베이스 구축을 위해 방문 사이트의 방문주소, 질의어, 사이트명, 카테고리 등과 같은 주요 스키마 속성들을 추출한다.
- 단계 2 : 속성 값이나 사례의 전처리 단계로 사례기반 추론에 적합하지 않는 내용이나 속성 값은 삭제 및 수정하여 추론 가능한 형태로 변형한다
- 단계 3 : 사례베이스 생성단계로 단계 1과 단계 2를 거치면서 처리된 사례들을 이용하여 사례베이스를 생성한다.

개인 사례베이스와 일반 사례베이스의 스킴(scheme) 구조는 동일하며 표 5.3과 같다. 사례기반 추론을 통한 정보 검색은 사용자가 입력한 질의어와 사례베이스의 사이트명, 카테고리, 설명부의 내용과 패턴 매칭을 실시하여 일치된 정보 중에서 4장에서 기술한 유사도 평가함수에 의해 결정된 카테고리 그룹에 속한 하부 정보들만 추출하여 관련 정보로 출력한다.

그림 5.16는 ‘연구실’이라는 질의어에 대해 사례기반 추론 검색을 실시한 결과 화면이다.

표 5.3 사례 베이스의 스킴 구조

URL	KEY WORD	TITLE	CATE	HIT
http://www.kadowa.com/	해양	해양 심층수 다목적 개발사업	과학, 학문/자연과학/지구과학/해양학	22
http://www.drillkorea.com/	해양	해양건드릴	비즈니스, 경제/업종별_회사/산업용품, 설비/재료, 소재/금속/주조, 금형, 가공/금형	18
http://www.oceanf.com/	해양	해양유선낚시	비즈니스, 경제/업종별_회사/레저, 야외활동/낚시/낚시배, 출조	48
http://oceanlove.com.ne.kr/	해양	이수호 해양개발연구소	과학, 학문/공학/조선, 해양공학	30
http://seafood.pknu.ac.kr/~marinbio/	해양	부경대학교 해양생물학과	과학, 학문/자연과학/생물학/해양생물학/학과, 연구실	26
http://www.kocean.org/	해양	김영구 - 해양법포럼	교육, 참고자료/대학교육/교수, 교직원	52
http://www.pusansek.com/	해양	해양소년단 부산탐험개척대	교육, 참고자료/초중고교육/기관, 단체/청소년단체/한국청소년연맹	61
http://plankton.cheju.ac.kr/	해양	제주대학교 해양플랑크톤 연구실	교육, 참고자료/대학교육/국공립대학교/제주대학교/연구실	12
http://rcoid.pknu.ac.kr/	해양	부경대학교 해양산업개발연구소	과학, 학문/자연과학/지구과학/해양학/연구소	17

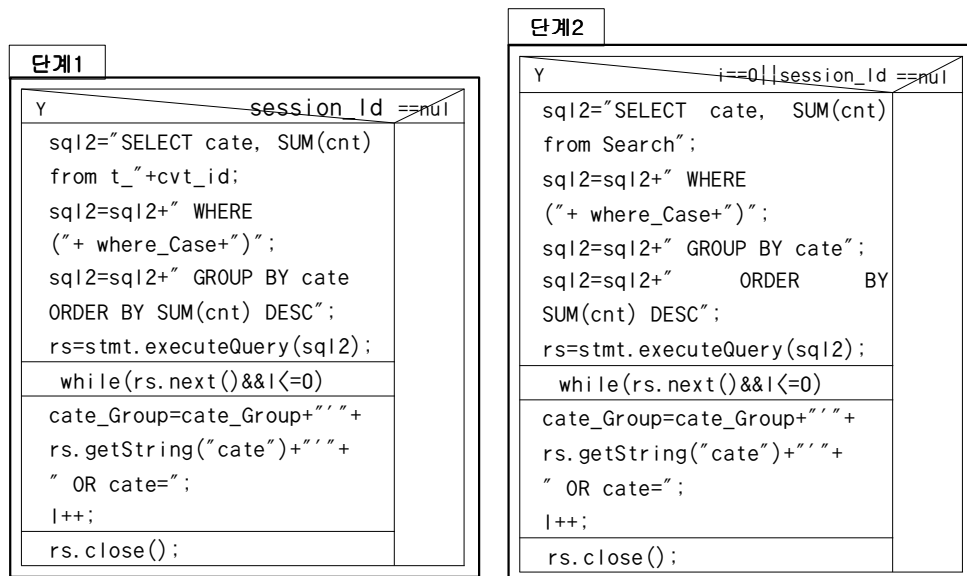


그림 5.15 유사도 평가 모듈의 구현 코드



그림 5.16 사례기반 추론 결과 화면

## 5.5 무선 인터넷 연결 모듈

### 5.5.1 유무선 통신의 접근 위상

무선 인터넷은 무선통신 기술의 발전으로 언제 어디서나 인터넷에 접속하는 이동성과 접근성을 통해 다양한 정보를 검색할 수 있는 환경을 제공하기 때문에 기존의 유선 인터넷의 시간적, 공간적 제약 사항을 극복할 수 있는 대안이 되어 가고 있다. 이러한 인터넷 사용 환경의 변화는 이동시에도 유선 인터넷 망의 정보 검색 기능을 무선 단말기에서도 정보의 손실 없이 같은 방법으로 사용하는 접근 기술을 요구하고 있다.<sup>[42]</sup>

현재 무선을 통한 유선 인터넷 정보 접근 위상은 그림 5.17과 같이 웹 서버와 무선 단말기 사이에 무선 인터넷 서비스 제공자를 통해 상호 연결되어 있으며 무선 가입자 망은 여러 기지국 및 교환기들을 통해 전화통화 연결과 같은 호 처리(call processing)과정을 거치게 된다.

그림 5.16의 실선 블록에서 MSC(mobile switching center)는 이동통신교환기로 호 처리 기능을 주로 담당하며 PSTN(public switched telephone network)망과의 접속을 제공한다. HLR(home location register)은 이동전화가입자의 기본 정보와 이동국의 위치정보, 각종 권한정보와 같은 부가 정보 등에 대한 관리를 담당하며 이동교환기가 호 연결을 하도록 지원하는 역할을 한다. IWF(inter-working function)는 망 연동장치로서 프로토콜 변환, 대역폭 변환, 서비스 품질(QoS : quality of service) 보장 등의 기능을 기본적으로 담당하고 특히 무선 이동 단말에서 인터넷 서비스를 제공하기 위해 ATM 기반의 무선 통신망과 인터넷간의 연동 기능을 지원한다.

무선 단말기는 브라우저의 종류에 따라 SMS폰, WAP폰, ME폰 등 여러 종류가 있고 단말기에서 사용되는 언어도 서비스 업체에 따라 SK-WML, UP-WML, mHTML, HDML 등 다양한 무선 마크업 언어가 있기 때문에 해당 단말기가 인식할 수 있는 언어로 정보를 제공하는 서버에만 접속해야 하는 문제점이 있다. 또한, 유무선 간의 프로토콜 차이로 인터넷 표준 언어인 HTML로 작성된 웹 문서는 무선 URL을 통해 접근하는 무선 단말기에서는 처리할 수 없는 문제점이 있다.

따라서 본 논문에서는 유선 인터넷 검색 엔진을 무선 단말기에서도 인식될 수 있도록 여러 다양한 무선 마크업 언어로 무선 웹사이트를 구축하여 무선 단말기에서도 지능형 정보 검색 기능을 제공하고자 한다.

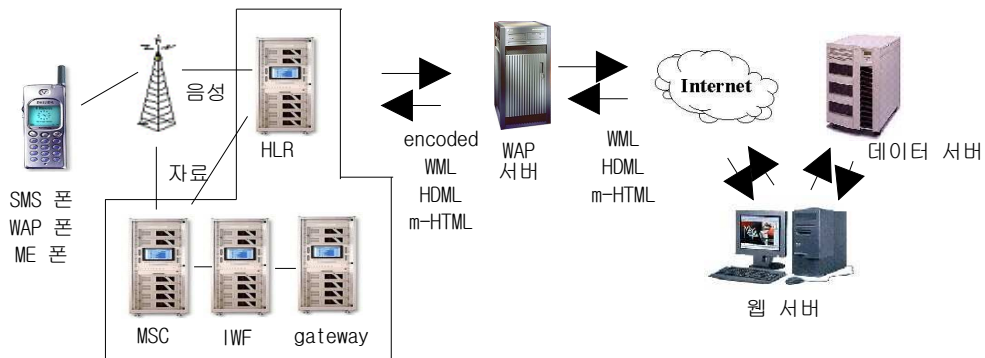


그림 5.17 무선통신을 통한 유선 인터넷 접근 위상

### 5.5.2 무선 인터넷 구현 기술

무선데이터서비스 사용자들이 무선 인터넷에 접근하기 위해서는 무선 브라우징 프로토콜이 필요하다. 무선 브라우징 프로토콜은 여러 종류가 있지만, 이동통신 단말기용으로 나온 HDML(handheld device markup language)언어를 사용하는 WAP(wireless application protocol) 방식이 기존 인터넷과 호환성에 중점을 두고 HTML을 근간으로 하는 마이크로소프트의 스팅거(stinger) 방식이나 일본 NTT의 I-mode 방식보다 빠르고 사용이 간편하여 사실상의 표준으로 간주되고 있다.

WAP 방식은 유선 인터넷에서 사용하는 HTTP 프로토콜에서 그 설계가 시작되었으며 유선 인터넷을 무선 통신에서도 이용할 수 있도록 하는데 그 목적을 두고 있다. WAP의 구조는 기본적으로 유선 인터넷의 클라이언트/서버 구조를 따르고 있지만 그림 5.18에서 나타나듯이 WAP방식은 기존의 TCP/IP 모델인 클라이언트/서버 사이에 HTML과 WML의 변환을 위해 WAP Proxy라 불리는 WAP 게이트웨이를 필요로 한다. 이 게이트웨이는 TCP/IP 네트워크 기반의 인터넷 데이터를 무선기반 네트워크에서 사용하는 바이너리 데이터로 인코딩, 디코딩 하는 역할을 수행한다. 즉, 웹서버에서 제공되는 기존의 유

선 인터넷 데이터를 무선 패킷으로 변경하여 무선 클라이언트에게 무선으로 전달하거나 역으로 무선 클라이언트로부터 WSP/WTP 형식으로 전달되는 데이터를 HTTP로 재 변환하여 유선 인터넷 선로를 통해 웹서버로 전달하는 기능을 담당한다. 일반적으로 웹서버에서 WML로 작성된 문서는 기존 캐릭터기반의 유선 데이터이며 서버에서 작성된 문서가 게이트웨이로 전달되면 게이트웨이는 전송 받은 WML을 캐릭터기반이 아닌 바이너리 기반으로 단말기에 전송한다. 또한 단말기에서 전송되어 오는 데이터를 다시 유선 기반의 문서로 만들어서 웹서버로 전송하는 과정을 거치게 된다. 무선 인터넷에서 문서 전달의 단계별 세부 처리 과정은 표 5.4에 나타내었다. WAP 프로토콜 구조는 그림 5.19의 음영 부분과 같이 5개의 계층으로 구성되어 있으며 각 계층을 구성하고 있는 프로토콜을 통칭하여 WAP 프로토콜 스택이라 부른다.

먼저 WAE(wireless application environment)는 어플리케이션이 동작할 수 있는 기본적인 계층으로 응용 어플리케이션의 개발 및 실행에 관계하는 모든 요소를 관장하는 계층이다. 즉, WAP 무선 인터넷 서비스 개발자에게 제한된 리소스를 가진 무선 단말기에서 사용 가능한 무선 인터넷 콘텐츠 개발에 필요한 사항들을 정의하고 있는 프로토콜이다.<sup>[43]</sup>

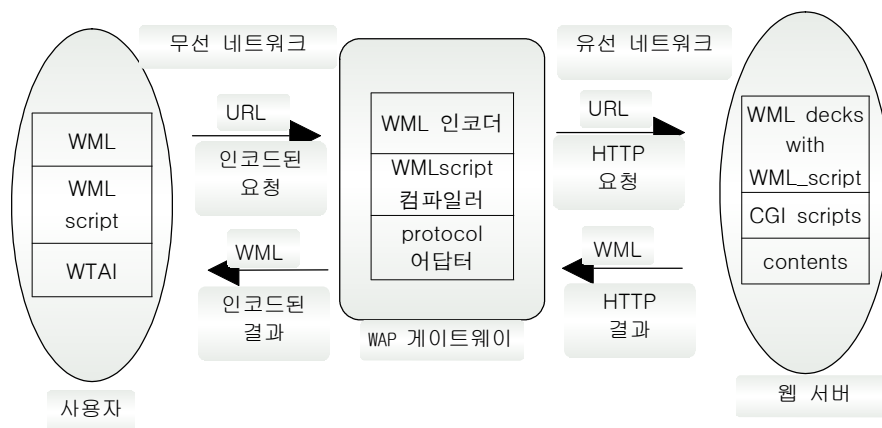


그림 5.18 WAP방식의 무선 클라이언트/서버 통신 모델

WSP(wireless session protocol)는 세션 서비스와 세션관리를 위한 정지 및 재개 기능을 제공하는 계층으로 WAP 프록시가 WSP 클라이언트를 표준 HTTP서버에 연결할 수 있도록 한다.

표 5.4 무선 인터넷의 단계별 처리 기능

단계	기능 및 역할
•단계 1	클라이언트는 특정 URL에 대한 접속을 요구한다
•단계 2	단말기 브라우저는 WSP로 게이트웨이에 연결하고 목적지 URL로 GET 응답을 전송한다.
•단계 3	게이트웨이는 URL에 명시된 웹 서버의 주소를 해석하고 해당 서버와 HTTP 세션을 구성 한다.
•단계 4	게이트웨이는 URL에 명시된 컨텐츠를 HTTP 프로토콜을 통해 웹 서버에 접속 한다.
•단계 5	웹 서버는 HTTP 프로토콜의 요구에 응답 한다.
•단계 6	게이트웨이는 웹서버로부터 받은 WML 소스를 WAE에서 규정된 해당 hexa 값으로 인코딩하여 클라이언트로 전송한다.
•단계 7	반대로 클라이언트의 응답은 게이트웨이로 역 경로를 설정하고 게이트웨이로 전달된 hexa 값을 WML로 디코딩하여 웹서버로 재 전송한다

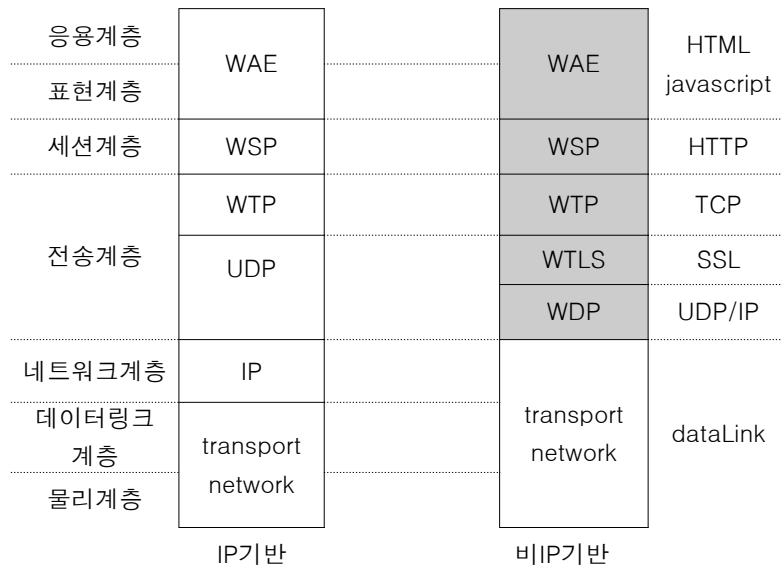


그림 5.19 WAP 프로토콜 스택 구조

WTP(wireless transaction protocol)는 무선 환경에 적합한 작은 트랜잭션 형태의 데이터 전송기능을 담당하는 계층으로 신뢰 및 비 신뢰성 전송과 오류 복구를 위한 재전송 기능을 제공한다. WTLS(wireless transport layer security)는 기존 인터넷의 TLS를 근간으로 작성된 보안 프로토콜이다. 인증, 부인봉쇄(non-repudiation), 무결성과 기밀성 등의 보안 서비스를 제공한다. WDP(wireless datagram protocol)는 점대점 전송을 위한 계층으로 다양한 네트워크 형태에 의해 지원되는 데이터를 운반하는 능력을 가진 전송 서비스 위에서 동작하며 상위계층 프로토콜에 일관된 서비스를 제공한다.

### 5.5.3 무선 에이전트의 구현 및 처리 절차

무선 에이전트의 구현은 무선 단말기의 물리적인 제약사항으로 인해 설계단계에서부터 디스플레이의 크기를 포함하여 네트워크 용량 및 메모리 자원 등을 고려하여야 한다. 특히 WML이 아직 표준화되지 않아 무선 통신 서비스 회사마다 조금씩 다른 언어환경으로 인터페이스 모듈을 구현하였다.

무선 에이전트의 인터페이스 모듈은 클라이언트 표현모듈과 데이터베이스 연동모듈로 구성된다. 표현모듈은 WAP 프로토콜 기반의 WML을 사용하였고 색인데이터베이스와의 연동모듈은 ASP를 사용하였다. 표현모듈은 하나의 덱(deck)으로 구성하고 하나의 덱은 여러 카드로 구성하였다. 덱은 WAP 응용서버에서 무선 단말기로 전송되는 WML의 최소단위이므로 사용자가 무선 인터넷 에이전트에 접속하면 사용자의 무선 단말기로 하나의 덱을 전송시킨다. 전송된 덱의 첫 번째 카드가 단말기에 표시되면 무선 단말기를 통해 데이터를 입력하거나 또는 다른 카드로 이동할 수 있도록 하였다. 무선 에이전트의 표현과 연동을 위한 인터페이스 모듈은 부록에 나타내었다.

그림 5.20은 무선 단말기에서 인터넷 정보를 검색하기 위한 모듈 구성도이다. 무선 단말기는 무선 전화망을 통해 무선 서비스제공자와 연결되어 있고,



검색엔진 서버와 이동통신 사업자는 WAP 게이트웨이를 통해 유선으로 연결된다. 이 게이트웨이는 무선 통신망을 통해 무선 단말기와 검색엔진 서버가 서로 동일한 환경에 접속되어 있는 것 처럼 보이기 위한 중계 기능을 담당한다. 따라서 무선 단말기에서 무선통신망을 통해 이동통신 사업자와 정상적으로 연결이 된다면 WAP 게이트웨이를 통해 검색 엔진 서버가 제공하는 HTTP 프로토콜 정보는 WAP 프로토콜로 변환이 되고 사용자는 무선 단말기에서 유선에서와 같은 방법으로 정보 검색을 할 수 있다.

그림 5.21은 무선 단말기에서 유선 검색에이전트에 접속하여 정보를 검색하는 과정의 흐름도이다. 먼저 무선망을 통해 무선단말기가 검색에이전트에 연결이 되면 무선 웹 서버는 무선 단말기의 하드웨어, 사용 언어, 프로토콜 등과 같은 특성을 파악한다. 이것은 검색 정보를 무선 단말기로 송신할 때 단말기의 특성에 따라 사용 언어를 달리해야 하기 때문이다. 무선 단말기의 특성은 HTTP 1.1 Protocol에 의해 정의된 User Agent Field 영역의 단말기 Header 정보를 분석하면 파악된다. 단말기의 특성이 파악된후 단말기의 특성에 맞게 작성된 시작 파일이 연결된다. 사용자는 정보 검색을 위해 질의어를 직접 입력하거나 정보 검색 시스템이 제공하는 메뉴를 통해 질의어를 입력한다.

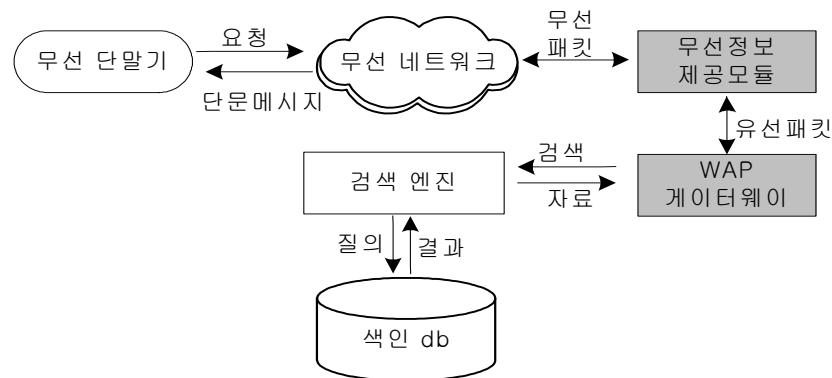


그림 5.20 무선 인터넷 망을 통한 검색에이전트 연동 모델

이때 무선 단말기 특성별로 작성된 WML은 WAP 게이트웨이를 거쳐 질의열(query string)형태로 검색 에이전트 서버에 전달한다. 전달된 질의열을 이용하여 검색 에이전트 서버는 색인데이터베이스에서 검색을 수행하고 질의 결과인 전체 레코드 셋을 무선 단말기의 특성에 따라 WML로 변환시켜 무선 단말기로 전송한다. 다시 무선 단말기에서는 전송 받은 전체 검색 결과 중 하나의 URL을 선택하면 해당 사이트에 접속하고 해당 서버에서 제공하는 카드 정보의 특성을 판단하게 된다. 만약 해당 단말기를 지원하는 WML로 작성된 웹 페이지일 경우에는 특별한 변환 작업 없이 송신할 수 있지만 다른 형식의 언어로 작성되었을 경우에는 변환 과정이 필요하다. 이러한 과정이 정상적으로 이행된다면 무선 웹 서버는 해당 사이트에서 선택한 페이지의 내용을 카드의 형태로 무선 단말기로 재 송신하여 유선 인터넷의 정보를 무선 단말기로 출력시킨다.

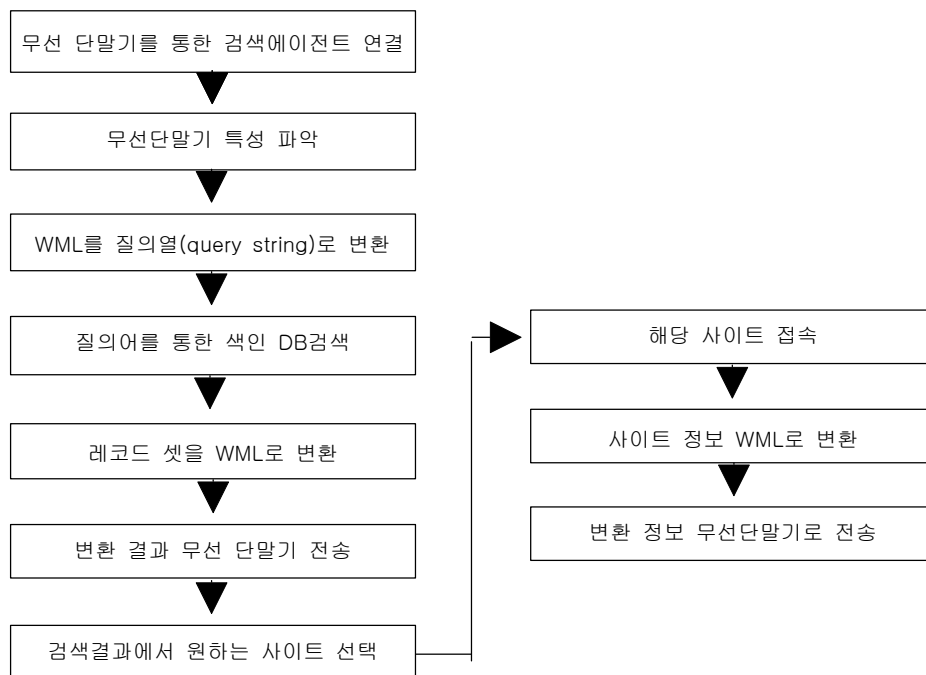


그림 5.21 무선 인터넷 정보 흐름도

#### 5.5.4 무선 인터넷의 접속 및 검토

접속 실험은 018 무선 단말기와 019 무선 단말기에서 각각 이루어졌다. 그림 5.22는 018 휴대폰에서 북마크를 통해 검색하는 화면이고 그림 5.23은 019 무선 단말기에서 URL을 직접 입력한 접속 화면이며 ‘대학’과 ‘해양’이라는 질의어에 대한 검색 내용을 차례로 보여주고 있다.

무선 단말기를 통한 검색 에이전트의 접속 실험에서는 무선단말기의 물리적인 제약사항, 즉 9,600bps 정도의 낮은 대역폭, 한번에 1,500 바이트 이하의 파일만 처리 할 수 있는 제한된 메모리와 기본 화면의 크기가 12글자 × 4줄의 작은 디스플레이 크기에 맞추어서 설계해야 하는 문제점으로 인해 유선에서와 같은 다양한 구현 기술은 발휘 할 수 없는 문제점이 있다. 특히 제한적인 사용자 입력환경으로 인해 독립된 지능적 추론 기능은 무선 단말기에서 구현하기 어렵고 유선 웹 서버에 의존적인 검색 기능을 수행할 수 밖에 없다.



그림 5.22 무선 단말기(018)에서의 검색 과정



그림 5.23 무선 단말기(019)에서의 검색 과정

## 제 6 장 실험 및 고찰

### 6.1 연관규칙 탐사 기법을 통한 실험

#### 6.1.1 실험성능 평가 방법

정보검색에서 성능평가는 일반적으로 검색된 결과의 정확율과 재현률의 측정으로 이루어진다. 정보검색 성능 평가의 척도로서 정확율과 재현률을 식 (6.1)과 식 (6.2)와 같이 정의하였다.

$$\text{정확율} = \frac{\text{주어진 질의어와 관련된 검색 문서 개수}}{\text{검색된 전체 문서들 개수}} \quad (6.1)$$

$$\text{재현률} = \frac{\text{주어진 질의어와 관련된 검색 문서 개수}}{\text{DB에 존재하는 관련된 모든 문서들 개수}} \quad (6.2)$$

본질적으로 어떠한 검색 알고리즘도 주어진 질의에 대해 자신의 문서 집합으로부터 관련된 모든 문서 결과를 항상 반환해 주지 못하고 또 항상 같은 정확도를 보장해 주지도 못한다. 즉 재현률이나 정확율은 주어진 질의어의 종류에 따라 다른 결과를 나타낼 수도 있다. 따라서 두 성능 평가율을 일정한 값으로 수렴시키기 위해서는 여러 카테고리 집합에서 균일한 분산을 갖도록 실험 데이터의 개수를 증가시킬 필요가 있다. 이것은 계수(計數)로 제공되는 실험 결과 값 자체보다는 동일한 질의어 그룹에 대한 검색방법의 상대적 비율이나 변화율에 더 많은 분석적 가치를 부여해야 함을 의미한다. 또한 정확율과 재현률의 가치 평가 기준은 사용자의 선호도나 사용 목적에 따라 그 중요도가 결정될 문제로서 직접 서로 비교할 수 없는 값들이다.

### 6.1.2 실험 환경 및 결과 분석

연관규칙 탐사 기법의 객관성을 보증하기 위해 무작위로 해양관련 주제 30개를 질의어로 추출하고 각 질의어를 6개씩 묶어 질의어 그룹으로 구성하였다. 또한 연관규칙 탐사 기법의 유효성을 평가하기 위해 연관추론 검색 외에 해양분야의 전문검색엔진의 검색, 범용검색엔진의 검색을 함께 수행하였다. 표 6.1은 질의 결과에 대한 정확율과 재현률에 대한 평균치이다.

표 6.1 연관규칙 탐사 기법의 검색방법에 따른 정확율과 재현률

그룹	연관규칙 탐사 검색		전문검색엔진 검색		범용검색엔진 검색
	정확율	재현률	정확율	재현률	정확율
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

범용검색엔진에서의 재현률을 나타내지 않은 이유는 검색에 이용된 범용검색엔진이 보유하는 데이터베이스의 관련 문서 개수를 산출할 수 없기 때문에 생략하였다. 실험에 따르면 재현률이 낮을수록 정확도는 높아지는 현상이 발생하였다. 이것은 재현률이 낮으면 추출되는 특정 색인어의 출현빈도가 높아지기 때문에 정확도는 상승한다. 즉 제한된 정보 영역에서의 색인어 추출은 특정 질의어를 검색하는데 더 작은 문서만을 참조하고 이것은 특정 색인어가 포함된 문서에서 질의어와 관련된 색인어간의 상호정보량의 편차가 줄어들어 정확율은 증가하는 것으로 분석할 수 있다.

질의어 신뢰도 평가는 질의어1과 질의어2간의 개별 연산에 대해서 질의어1에 주어진 항 S1을 기준으로 질의어2의 항인 S2, S3, S4, ..., Sn 간의 상호 교차 평가를 통해 연관규칙을 선정한다. 예를 들어 최소 신뢰도로 임계값 0.3이

주어진 경우 그림 6.1과 같이 항S1과 항S2, 항S3, 항S4와의 불린 연산이 이루어졌을 때 연관규칙 결정을 위한 신뢰도의 변이값을 구할 수 있다. 표 6.2는 그림 6.1의 연관규칙 신뢰도 변이값을 나타내었고 그림 6.2는 표 6.2를 도식화한 것이다. 이때 신뢰도의 변화 추이는 질의어 상호간의 연산 결과에 따라 계속 변화하며 같은 질의어에 대한 불린 연산의 횟수가 증가할수록 질의어들간의 확신도는 증가하고 있음을 보여주고 있다.

표 6.2 질의어1(항S1)을 기준으로 한 신뢰도 평가

연산		항2		S2		S3		S4	
		신뢰도	채택	신뢰도	채택	신뢰도	채택		
초기 상태	$S1 \wedge S2, S1 \wedge S3, S1 \vee S4$	33.3%	√	33.3%	√	16.6%	×		
실행1	$S1 \vee S2$	37.5%	√	25%	×	12.5%	×		
실행2	$S1 \wedge S3$	30%	√	40%	√	10%	×		
실행3	$S1 \vee S3$	25%	×	41.7%	√	8.3%	×		

횟수	불린 연산	상태
1	검색어: 해양 AND 검색어2: 유람선	초기 상태
2	검색어: 해양 AND 검색어2: 요트	
3	검색어: 해양 OR 검색어2: 박용	
4	검색어: 해양 OR 검색어2: 유람선	실행1
5	검색어: 해양 AND 검색어2: 요트	실행2
6	검색어: 해양 OR 검색어2: 요트	실행3

그림 6.1 연관규칙 결정을 위한 두 개의 질의어의 불린 연산

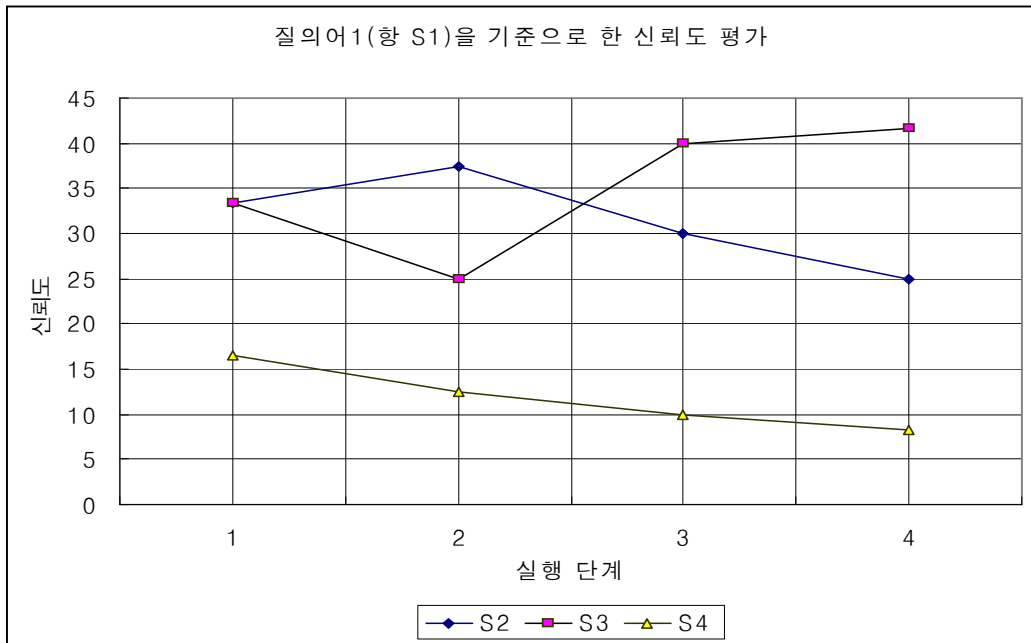


그림 6.2 표 6.2에 대한 신뢰도 변이 그래프

연관추론의 연관성 실험에서는 특정 분야의 전문지식이 부족한 일반 사용자가 하나의 질의어만 입력해도 그와 관련된 다른 연관어(associated word)를 함께 보여줌으로써 보다 폭 넓은 정보를 사용자에게 제공할 수 있다. 또한 신뢰도는 초기에 동적으로 변화하다가 횡수가 증가할수록 항들 사이에 일정한 패턴이 유지되어 연관규칙이 생성된 것을 확인할 수 있고 다른 검색방법에 대한 정확율과 재현률의 비교 실험에서는 연관규칙 탐사 기법이 적용된 검색 에이전트가 평균적으로 불 때 전문검색엔진에 비해 재현률은 높고 범용검색엔진에 비해서는 정확율이 높은 것으로 나타났다.

## 6.2 사례기반 추론 기법을 통한 실험

본 논문에서는 사례기반 추론의 정확성과 관련문서 수를 평가하기 위해 해



양정보를 저장하고 있는 로컬 색인 데이터베이스로부터 수집한 7,760개의 문서 자료를 이용하였다. 이 문서들은 전체 12개의 대분류 카테고리 분류되어 있고 대분류 카테고리 밑에 최고 7개의 하부 카테고리 그룹 레벨을 가지고 있다. 실험에 사용된 검색어 개수는 총 30개이며 검색 카테고리 그룹의 개수는 1,911개이다.

본 논문에서의 실험은 그림 6.3과 같이 네 가지 검색방법으로 수행된다. 첫 번째는 사례기반검색으로 사용자별 개인사례정보를 바탕으로 유사도가 가장 높은 카테고리 그룹의 정보만을 제공하는 경우이다. 두 번째로 AI-SEA일반검색은 해양 관련 내용으로 필터된 색인데이터베이스 정보들을 대상으로 검색을 수행하는 과정이며, 세 번째는 전문웹검색으로 범용검색엔진에서 카테고리별로 분류된 정보를 대상으로 검색을 수행한 경우이다. 네 번째는 웹페이지 검색으로 범용검색엔진으로 일반 웹 문서 전체를 대상으로 검색을 수행하는 경우이다.

실험은 네 가지 검색방법에 대해 해양관련 주제로 30개의 질의어를 무작위로 선택하고 관련 수준별로 5개씩 그룹화하고 검색하여 각 그룹별 질의 결과에 대한 평균치를 정확율과 재현률로 표 6.3에 나타내었다. 전문 웹검색과 웹페이지 검색에서 재현률을 나타낼 수 없는 이유는 검색에 이용된 범용검색엔진에서 보유하는 데이터베이스의 관련 문서 개수를 산출할 수 없기 때문이다.

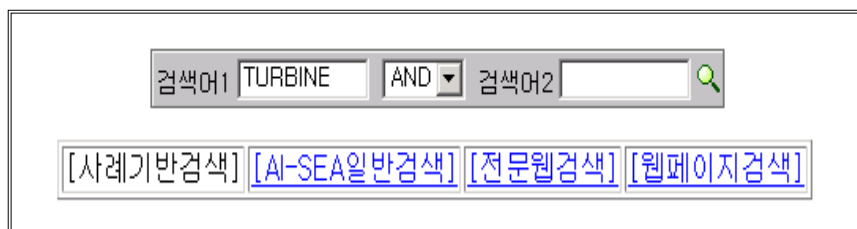


그림 6.3 질의어 입력 창과 검색 종류

표 6.3 사례기반 추론 기법의 검색방법에 따른 정확율과 재현률

그룹	사례기반 검색		전문검색엔진 검색		전문웹검색	웹페이지검색
	정확율	재현률	정확율	재현률	정확율	정확율
5	0.96	0.76	0.82	0.93	0.74	0.67
10	0.94	0.77	0.82	0.92	0.72	0.63
15	0.95	0.74	0.81	0.94	0.71	0.65
20	0.98	0.55	0.78	0.96	0.70	0.75
25	0.97	0.72	0.80	0.95	0.87	0.85
30	0.96	0.73	0.81	0.93	0.25	0.12

그림 6.4는 표 6.3의 자료 중 사례기반검색의 관련 수준에 따른 정확율과 재현률의 변화추이를 그래프로 나타내었다. 그래프 결과에 따르면 일반적으로 정확율과 재현률은 상대적으로 역비례 관계를 가지는 것으로 나타났다. 정확율은 미리 필터된 영역에서 자료를 추출하기 때문에 고른 편차를 보인 반면 재현률은 특정한 관련 수준에서 변화폭이 큰 경우가 있는데 이것은 학습된 자료의 양이 충분하지 않거나 개인적 검색 성향이 일반인들의 검색성향과 다른 패턴을 가질 경우 다른 관련 레벨보다 월등히 낮은 관련성을 갖는 검색 결과를 나타내는 것으로 보인다.

그림 6.5는 표 6.3의 자료 중 전문검색엔진에서의 정확율과 재현률의 변화추이를 그래프로 나타내었다. 전문검색엔진에서의 검색 방법은 색인데이터베이스에 저장된 정보가 해양관련 정보로 한정된 전문검색이기 때문에 정확율과 재현률의 변화 추이(推移)는 큰 차이를 보이지 않고 이 둘의 상관 관계도 사례검색과 같이 상대적으로 역비례 관계를 가지는 것으로 나타났다.

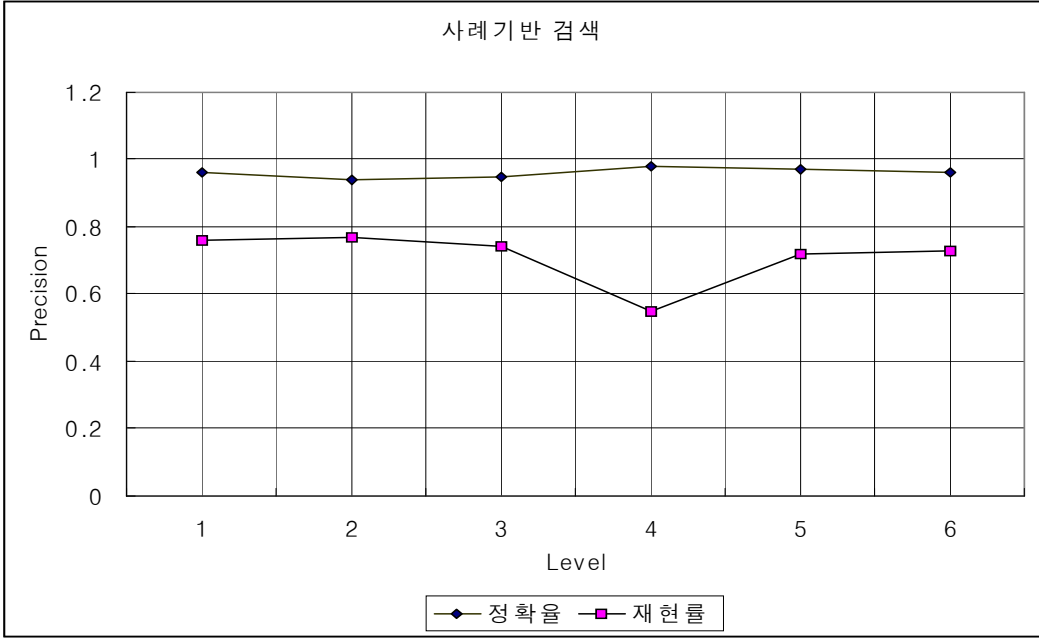


그림 6.4 사례기반검색에서 정확율과 재현률의 변화 추이

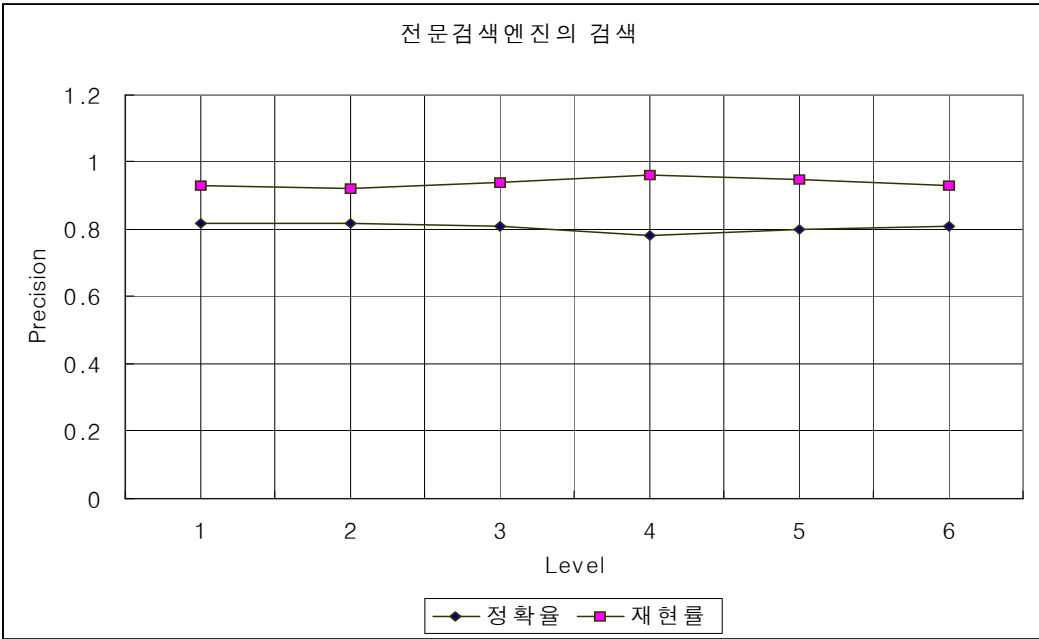


그림 6.5 전문검색엔진을 통한 정확율과 재현률의 변화 추이

그림 6.6은 표 6.3의 자료 중 사례기반 검색과 전문검색엔진 검색에서 정확율과 재현률의 변화 추이를 보여 주고 있다. 그래프 결과에 따르면 대체로 정확율과 재현률은 상대적 역비례 관계를 가지는 것으로 나타났다. 또한 사례기반검색이 전문검색엔진 검색보다 정확율이 높고 재현률은 전문검색엔진 검색이 사례기반검색에 비해 평균적으로 높은 것으로 나타났다. 이것은 개인의 검색성향을 고려한 사례기반검색이 한정된 카테고리 그룹 정보를 집중적으로 추출하기 때문인 것으로 보인다.

그림 6.7은 범용 검색엔진에서 전문웹검색과 웹페이지 검색의 정확율을 비교하고 있다. 평균적으로 전문웹검색이 일반 문서전체를 대상으로 하는 웹페이지 검색에 비해 상대적으로 높은 정확율을 나타내고 있다. 이것은 전문웹검색이 주제별로 분류된 영역에서 검색을 수행하기 때문이다. 이 그림에서 정확율이 큰 폭으로 감소하고 있는 경우가 있는데 이것은 특정한 질의어그룹에서 예를 들어 조선(造船)이라는 질의어를 검색할 경우 조선(朝鮮)이나 조선일보(朝鮮日報)의 웹 문서가 추출되는 경우에는 사용자의 의도와는 완전히 다른 결과를 초래하기 때문에 정확율을 현저히 낮출 가능성이 있다.

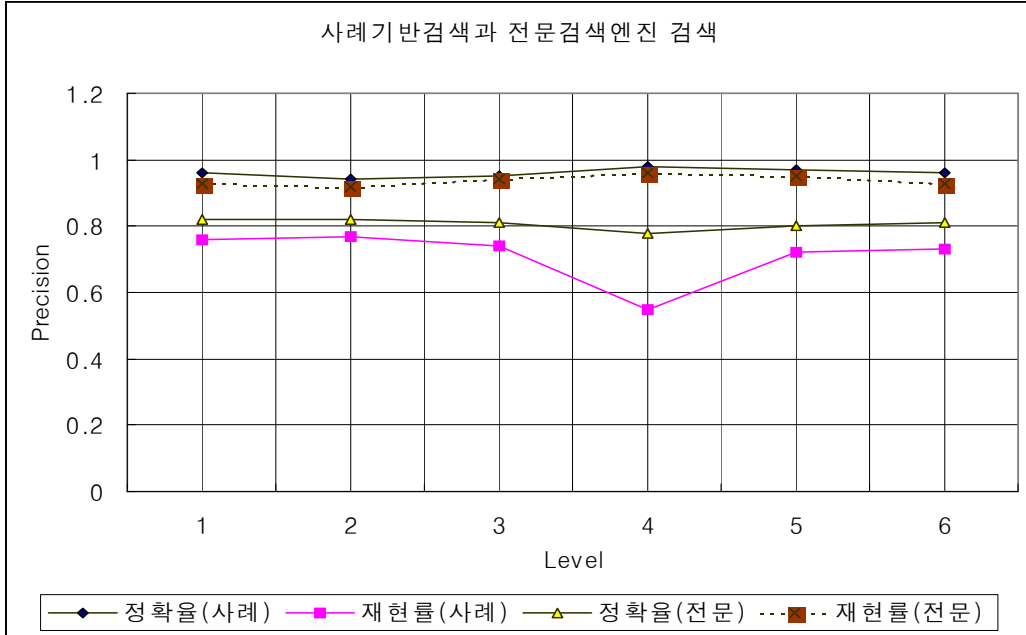


그림 6.6 사례기반검색과 전문검색의 정확율과 재현률의 변화추이

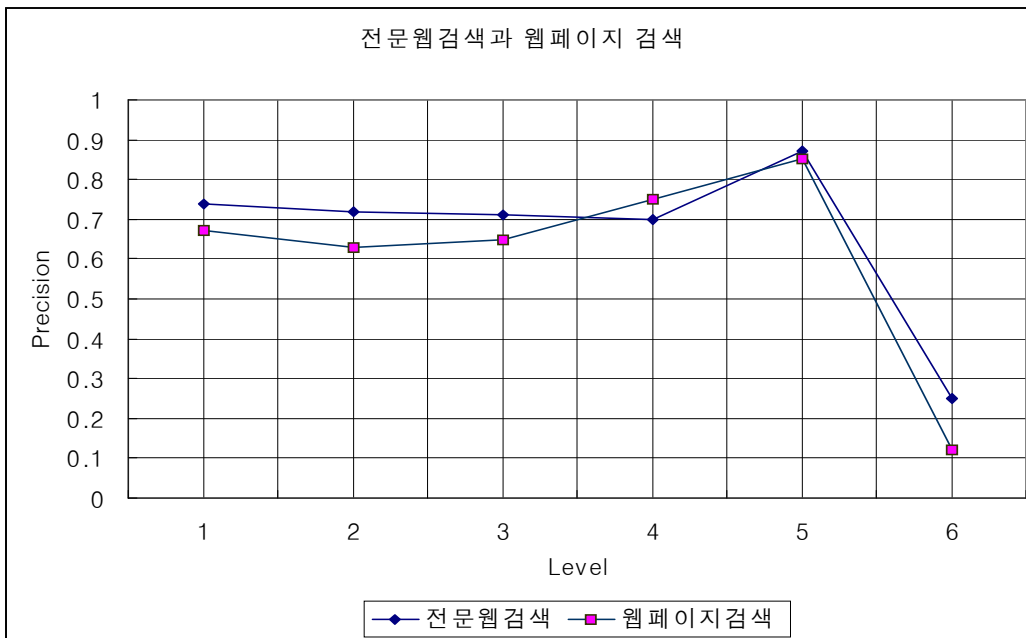


그림 6.7 전문웹검색과 웹페이지 검색에서의 정확율 변화 추이

## 제 7 장 결 론

본 논문에서는 개인화되고 연관성 있는 정보를 사용자에게 제공하기 위해 연관규칙 탐사 기법과 사례기반 추론 기법을 웹 검색엔진에 응용하여 유무선 통합 지능형 검색 에이전트 시스템을 구현하였다. 연관규칙 탐사 기법을 위해서는 ‘그룹화 규칙 생성 알고리즘’을 개발하였고 사례기반 추론 기법의 구현을 위해서는 ‘인지적 확률모델기반의 사례추론 알고리즘’을 개발하였다.

‘그룹화 규칙 생성 알고리즘’은 지지도와 신뢰도를 계산하여 연관자료의 확신도를 측정하고 ‘인지적 확률모델기반의 사례추론 알고리즘’은 유사도를 측정하여 관련성 정도를 평가하고 있다. 이러한 알고리즘은 검색 방법에 있어서 기존의 알고리즘에 비해 사용자의 검색 의도를 의미론적으로 해석할 수 있고 자료의 재현률과 정확율을 개선할 수 있다.

유사도 측정은 통계모델의 일종인 ‘유사 범주화 알고리즘’을 개발하여 평가 함수를 구성했다. 이 함수는 사이트 방문 회수를 카테고리별로 카운트하고 카테고리 그룹을 순서화 하여 그 카테고리 그룹에 속하는 하부 트랜잭션들을 사례 정보들로 제공한다. 이 알고리즘에 의해 질의어의 의미론적 해석을 담보하고 유형분석의 검색 효율을 높였다.

연관규칙 탐사 기법을 통해 특정 분야의 전문지식이 부족한 일반 사용자가 하나의 질의어만 입력해도 그와 관련된 다른 연관어도 함께 제공함으로써 보다 연관성 높은 정보를 제공할 수 있고 사례기반 추론 기법을 통해서도 개인 사용자나 일반 사용자의 검색 성향을 사례로 축적하고 유사도를 측정하여 사용자의 선호도에 따라 개인화된 검색기능을 제공한다.

사용자 모델링을 통한 학습은 사례베이스 관리 모듈이 유사사례추출과 적응 단계에서 여과된 정보를 사례표현 단계로 적절히 피드백하고 사용자의 카테고리 그룹을 지속적으로 변경시켜 지식의 확장과 적응기능을 가진다.

또한 유선 인터넷 정보의 무선 접근 방법에 대한 연구에서는 단말기의 특성에 투명하고 독립적으로 접근하는 WAP 기술을 웹 검색 에이전트에 응용하여 정보의 접근성 및 이동성을 높일 수 있다는 점을 확인하였다.

제안된 연관규칙 탐사와 사례기반 추론을 통한 정확율과 재현률에 대한 평가 실험에서는 평균적으로 정확율과 재현률은 역비례 관계를 가지는 것으로 확인되었으며 연관규칙탐사 검색에서는 신뢰도가 실행 횟수에 따라 동적으로 계속 변화하므로 보다 정확한 연관규칙이 검색에 반영될 수 있었다. 또한 정확율과 재현률의 비교 실험에서는 연관규칙탐사 검색이 전문검색엔진에 비해 재현률은 높고 범용검색엔진에 비해서는 정확율이 높은 것으로 나타났다. 사례기반 검색과 전문 검색엔진의 비교실험에서는 사례기반 검색이 전문검색엔진 보다 평균적으로 정확율이 높고 사례기반 검색과 범용검색엔진과의 비교 실험에서는 사례기반 검색이 범용검색엔진보다 정확율이 높은 것으로 나타났다.

이것은 사용자의 검색성향을 통계적으로 고려한 연관규칙 탐사 및 사례기반 추론 검색이 정확율 및 재현률에서 다른 검색 엔진에 비해 우수한 결과를 가져올 수 있음을 입증한 것으로 사료된다.

## 참고문헌

- [1] 김준태, 유건아, “인터넷 정보검색 시스템의 연구동향”, 전기학회지 제48권 제3호, pp. 52-59, 1999.
- [2] 하창승, 류길수, “사례기반 추론을 이용한 지능형 웹 검색 에이전트의 설계 및 구현”, 한국컴퓨터정보학회논문지 제8권 1호, pp. 20-29, 2003.
- [3] Harter, Stephen P. “A probabilistic approach to automatic keyword indexing: part II. an algorithm for probabilistic indexing”, *Journal of the American Society for Information Science*, vol. 26, no. 5, pp. 280-289, 1975.
- [4] Bruce Krulwich and Chad Burkey, “The InfoFinder Agent: Learning User Interests through Heuris”, *IEEE Expert/Intelligent System and Their Application*, vol. 12 no. 5, 1997.
- [5] <http://www.koreaweekly.co.kr/netInfo/200102/ni20010206192215n001159.htm>
- [6] 하창승, 윤병수, 류길수, “연관규칙탐사기법을 이용한 해양전문검색에진에서의 질의어 처리에 관한 연구”, 한국컴퓨터정보학회논문지 제8권 2호, pp. 8-15, 2003.
- [7] 전진욱, 배인환, “정보 수집 에이전트를 사용한 어린이 교육 정보 검색 시스템의 설계 및 구현”, 한국인터넷정보학회 제3권 2호, pp. 97-108, 2002.
- [8] M. Bauer, D. Dengler, and G. Paul, “Instructible Information Agents for Web Mining”, *In Proceedings of the 2000 International Conference on Intelligent User Interfaces*, pp. 21-28, 2000.
- [9] Salton and M. J. McGill, *An Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.
- [10] T. Joachims, D. Freitag, and T. Mitchell, “WebWatcher: A tour guide



- for the WWW”, *In Proceedings of IJCAI-97*, 1997.
- [11] H. Lieberman, “Letizia: An agent that assists web browsing”, *In Proceedings of IJCAI-95*, pp. 475-480, 1995.
- [12] <http://www.wisewire.com>
- [13] Rocchio, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313-323, Prentice-Hall, 1971.
- [14] L. Gravano, H. Garcia-Molina, and A. Tomastic, “The Effectiveness of GLOSS for the Text-Database Discovery Problem”, *In Proceedings of ACM SIGMOD*, 1994.
- [15] Adele Howe and Daniedl Dreilinger, “SavvySearch: A Meta-Search Engine that Learns Which Search Engines to Query”, *AI Magazine*, vol. 18 no. 2, Summer 1997.
- [16] Alexandrous Moukas and Giorgos Zacharia, “Evolving a multi-agent information filtering solution in Amalthea”, *In Proceedings of the 1st International Conference on Autonomous Agents*, 1997.
- [17] Clinton Wong, *Web Client Programming with Perl*, O'REILLY, March 1997.
- [18] 맹성현, 주종철, “문서구조화와 정보검색”, 정보과학회논문지, 한국정보과학회, 제16권 8호, pp. 6-14, 1998.
- [19] C. Faloutsos and S. Christodoulakis, “Signature Files: An Access Method for Documents and Its Analytical Performance Evaluation”, *ACM Trans. on Office Information System*, vol. 2 no. 4, pp. 267-288, 1984.
- [20] Thorsten Jachims, “A Probabilistic Analysis of the Rocchio Algorithm with TF-IDF for Text Categorization”, *In Proceedings 14th*

- International Conference on Machine Learning*, pp. 143 - 151, 1997.
- [21] 이종혁, 정용근, 남기곤, 윤태훈, 김재창, 박의열, 이양성, “G 연산자의 신호분석 특성을 이용한 음성인식 신경회로망에 관한연구”, 전자공학회논문지, 제29권 B편, 10호, pp. 90-98, 1992.
- [22] 合原一辛 편저, 정호선, 여진경 공역, *뇌전위와 카오스*, Ohm사. 1994.
- [23] 정태진, “강화 학습을 이용한 웹 정보 검색”, 서울대학교 석사논문, 2002.
- [24] 김상범 등, “문서범주화를 위한 선형 분류기와 kNN의 결합 모델”, 한국인지과학회 춘계학술대회 논문집. pp. 225-231, 1999.
- [25] P. Nordin and W. Banzhaf, “Real Time Control of a khepera Robot using Genetic Programming”, *Cybernetics and Control*, vol. 26, no. 3, pp. 533-561, 1997.
- [26] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.
- [27] J. R. Quinlan, “Simplifying decision trees”, *International Journal of Man-Machine Studies*, pp. 221-234, 1987.
- [28] G. Salton, *The SMART Retrieval System-Experiments in Automatic Document Processing*, Prentice-Hall, Inc., Englewood Cliffs NJ, 1971.
- [29] L. Gravano and H. Garcia-Molina, “Generalizing GLOSS to Vector-Space Databases and Broker Hierarchies”, *In Proceedings of VLDB*, 1995.
- [30] 최용석, “분산된 웹 데이터베이스에서의 정보검색 신경망 에이전트”, 서울대학교 박사학위논문, 2000.
- [31] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, *Knowledge Discovery in Databases : An Overview*, AAAI Press, 1991.
- [32] <http://ai.ewha.ac.kr/~hyeyeon/mining/웹마이닝.pdf>
- [33] 허명희, “이중 K-평균군집화”, 응용통계연구, 제13권 2호, pp. 343-352,

- 2000.
- [34] 조동영, “SuffixSpan : 순차패턴마이닝을 위한 형식적 접근방법”, 한국컴퓨터교육학회 제5권 4호, pp. 53-61, 2000.
  - [35] J.Borges and M.Levine, “A Fine Grained Heuristic to Caputre Web Navigation Pattern”, ACM SIGKDD 2000 Conference, vol. 2, Issue 1, pp. 40-50, 2000.
  - [36] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules in Large Databases”, In Proceedings of ACMOD Conference on Management of Data, Washington D.C., pp. 207-216, 1993.
  - [37] [http://cs.sungshin.ac.kr/~jpark/HOME/What\\_is\\_DM/kiss\\_review.pdf](http://cs.sungshin.ac.kr/~jpark/HOME/What_is_DM/kiss_review.pdf)
  - [38] Watson, I., “Case-Based Reasoning is a Methodology not a Technology”, *Knowledge Based Systems*, pp. 303-308, 1999.
  - [39] Schank, R. C., “Identification and conceptualizations underlying natural language”, *Computer Models of Thought and Language*, 1973.
  - [40] J. Xiong, “Case-Based Reasoning in Forecast Insect Pests”, *In Proceedings of PACES 95*, pp. 627-630, 1995.
  - [41] <http://lis.yonsei.ac.kr/course/under/ir2002/clustering.ppt>
  - [42] 이봉규, “XML기반의 Mobile기술에 관한 연구”, 정보처리학회, 제8권 3호, 2001.
  - [43] Steve Mann and Scott Sbihli, *The Wireless Application Protocol, A Wiley Tech Brief*, 2000.

## 부 록

### ■ 무선 에이전트의 인터페이스 모듈 코드

```
<%Response.ContentType="text/vnd.wap.wml"%>
<%
    Response.Expires = -1
    Response.AddHeader "Pragma", "no-cache"
    Response.AddHeader "cache-control", "no-store"
    search = request("search")
%>
<?xml version="1.0" encoding="KS_C_5601-1987"?>
<!DOCTYPE wml PUBLIC "-//PHONE.COM//DTD WML 1.1//EN"
    "http://www.phone.com/dtd/wml11.dtd">
<wml>
    <card ordered="true">
        <onevent type="onenterforward">
            <refresh><setvar name="SearchString" value=""/>
            </refresh></onevent>
            <input name="SearchString"/>
            <do type="accept" label="menu">
                <go href="#menu"/>
            </do>
        </card>
        <card id="menu">
            <select>
                <option
                    onpick="result.asp?search=<%=search%>&amp;SearchString=
                    ${SearchString}" title="select">search</option>
            </select>
        </card>
    </wml>
```

## 감사의 글

논문의 마지막 장인 감사의 글을 쓰면서 이 논문이 있기까지 도와 주신 많은 분들을 한분씩 생각합니다. 13년전 석사과정의 첫 제자로 선뜻 저를 받아 주시고 박사학위를 받을 때까지 관심과 후원을 아끼지 않으신 류길수 교수님께 무한한 감사의 마음을 드립니다. 언제나 당신을 제 마음의 큰 나무로 여기며 살아가겠습니다.

석사학위와 박사학위 심사에서 심사위원과 심사위원장을 맡으시며 격려와 용기를 주시고 언제나 따뜻한 미소를 잃지 않으신 이상배 교수님, 논문의 마지막 토씨 하나까지 꼼꼼히 검토해 주시고 자신의 경험과 이상을 심어 주시며 거울 앞에 선 제 모습을 보는 것같아 형이라 부르고 싶은 조석제 교수님, 해박한 지식과 경륜으로 논문의 흐름과 방향을 짚어주시고 지도해 주신 동명정보대학교의 김성진 교수님, 논문 심사 내내 균형과 여유, 조화로움을 함께 보여주신 기계연구원의 정경열 박사님에게 깊은 감사를 드립니다.

박사학위 취득을 저보다 더 기뻐해 주신 마음의 동지인 김종철교수님, 안말숙 교수님, 또한 힘들었던 지난 학기 내내 같은 고행의 길을 함께 걸어오며 서로 격려하고 정을 나누었던 정정수 교수님께도 감사의 마음을 드립니다.

실험과 논문 작성에 많은 도움을 주신 조준모 교수님, 윤병수 후배, 박은정, 정재열과 일일이 거명하지는 못하지만 마음의 후원자인 인공지능연구실의 여러 선후배에게도 감사의 마음을 전합니다. 또한 바쁜 입시 기간 중에도 많은 배려와 시간을 할애해주신 우리 학과의 여러 교수님들께도 이 지면을 빌어 감사의 말씀을 드립니다.

학위취득에 작은 영광이라도 있다면 그것은 오로지 가족의 후원과 인내에 있었음을 알고 있습니다. 논문을 핑계로 늦은 귀가를 언제나 묵묵히 참아준 아이들과 아내에게 미안하다는 말과 이 자식에게 기대와 희망을 걸고 살아오

신 어머님께 영광을 돌립니다.

이제까지 학문의 작은 강줄기를 따라오면서 학위가 학문의 끝이라 생각했지만 학해(學海)에 이르러 바라보는 더 넓은 학문의 바다를 보매 다시 한번 부끄러움과 두려움을 갖습니다. 지금과 같은 마음으로 생의 마지막까지 학문에 매진하며 살아갔으면 좋겠습니다.

계미년 세모(歲暮) 용당에서 ...