韓國 컴퓨터情報學會 論文誌
第8卷 第1號, 2003. 3.

2003-8-1-1-4

사례기반 추론을 이용한 지능형 웹 검색 에이전트의 설계 및 구현

하 창 승* 류 길 수**

Design and Implementation of Intelligent Web Search Agent using Case Based Reasoning

Chang-seung Ha* Keel-soo Rhyu**

요 약

웹에서 정보의 양이 급속히 증대됨에 따라 자신에게 맞는 정보를 찾는데 더 많은 시간을 투자하고 있다. 이러한 문제를 해결하기 위해서는 검색에이전트가 사용자의 선호도나 검색 목적에 따라 개인화된 검색기능을 제공하여야한다. 따라서 검색에이전트가 이러한 기능을 제공하기 위해 본 연구에서는 사용자가 과거에 검색과 관련된 경험적 지식을 축적하고 이 지식을 이용하여 새로운 질의어가 주어졌을 때 가장 관련성이 높은 카테고리 그룹을 결정하는 유사도 평가 방법을 통해 각 개인의 검색성향을 통계적으로 고려한 사례기반 추론기법을 제안한다. 사례기반 추론기법과 다른 일반검색 방법이 함께 적용된 검색엔진에서 실시한 성능 평가는 사례기반 추론기법이 일반 검색 방법에 비해 정확률에서 우수한 결과를 보였다.

Abstract

According as quantity of information is augmented rapidly in World Wide Web, users are investing more times finding correct information to own. Search function that a search agent is personalized according to user's preference degree or search objective to solve these problem should be offered. Therefore, a search agent accumulates experienced knowledge connected with user's past search in this research. When new query was given, search agent offered learning function of intelligence that decides category group through estimation method of similarity using this knowledge. So this paper showed that case based search can bring superior result in the correctness rate than other search method.

* 동명대학 정보통신계열 조교수
** 한국해양대학교 기계정보공학부 교수

접수일자 : 2003. 1.13
심사완료 : 2003. 3. 5

Ⅰ. 서론

최근 인터넷에서 획득할 수 있는 정보의 양이 급속히 증대됨에 따라 자신에게 유용한 정보를 찾는데 점점 더 많은 시간을 투자해야만 한다. 즉, 현재 웹 상에서 색인 가능한 문서의 수가 25억 개를 넘었고, 동적으로 생성되는 웹 페이지의 수는 5500억개 정도이며, 하루에 새롭게 생겨나는 웹 문서도 백 만개를 넘어서고 있어 검색엔진의 처리 시간을 지연시키고 있다. 또한 문서의 양적 증가로 인한 처리지연 이외에도 웹 문서의 분류나 표현규칙이 아직 표준화되어 있지 않아 특정 주제에 대한 사용자의 정보 요구를 정확하게 인식하지 못하고 있거나 사용자의 선호도나 검색 목적에 따라 개인화된 검색기능을 제공하지 못하는 질적인 문제점도 있다[1].

기존 범용검색엔진의 근본적인 문제점은 급증하는 웹 문서를 수용하기 위해 너비우선탐색법과 같은 단순한 방법으로 링크 그래프를 방문하기 때문에 검색 효율이 낮고 반복적으로 동작하는 검색 로봇(crawler)은 인터넷 트래픽을 증가시키고 있다. 정보의 분류에 있어서도 동적으로 변하는 웹의 성질 때문에 색인된 정보를 다시 검색해서 변화된 문서를 반영하는 검색 주기(reference rate)도 증가하는 정보의 량을 반영하기 힘들 정도이다. 이러한 문제는 특정 주제와 관련 없는 분야까지 검색하여 응답시간을 저해하는 문제를 낳기도 한다. 또한 문서의 의미를 분석하여 카테고리 그룹을 분류하거나 특정한 응용영역(specified domain)에 대한 경험적 지식을 재활용하는 학습 기능도 제공되지 않았다. 따라서 거대한 가상의 지식공간을 대상으로 하는 정보검색에서는 신속한 검색이나 풍부한 자료의 제공 못지 않게 사용자의 의도를 정확히 파악하여 사용자별로 개별화된 전문 지식을 제공할 수 있는 검색엔진과 개별화된 정보를 제공하기 위해 문제 영역 지식을 이용하거나 사용자의 선호도를 고려하는 지능적 검색 에이전트에 대한 연구가 필요하다.

지능적 검색 에이전트는 현재 검색을 요청하는 사용자가 누구인가에 따라서 사용자의 취향에 따른 검색결과를 제공할 수 있어야 한다. 정보 검색 에이전트가 지능적 학습 능력을 가지지 못한다면 질의에 대해 아무리 풍부한 관련 문서를 제공할 수 있다고 하더라도 사용자의 취향에 맞지 않는 결과들로써 사용자의 불편만 가중시킨다. 이러한 문제를 해결하기 위해 정보 검색 에이전트는 사용자의 과거 사례나 경험을 기억하였다가 이를 새로운 작업수행에 이용하는 학습과정이 필요하며 학습과정은 사용자에 따라 개별화된 단계까지 정보를 모델링시켜야 한다[2]. 그러므로 사용자 개인별로 사례기반학습과 같은 기계학습 방식을 활용하여 사용자별로 독립된 프로파일을 구성하고 이것을 검색에 이용할 수 있어야 한다. 따라서 본 연구에서는 사용자의 과거 검색과 관련된 경험적 지식을 축적하고 이 지식을 이용하여 새로운 질의어가 주어졌을 때 가장 관련성이 높은 카테고리 그룹을 결정하는 유사도 평가 방법을 통해 지능적 학습 기능을 제공하고자 한다.

Ⅱ. 지능적 웹 검색을 위한 이론적 연구

1. 정보 검색 에이전트의 고찰

현재 분산된 웹 환경에서 일반적으로 사용되는 정보검색은 사용자의 질의를 웹 상의 여러 데이터베이스들에 동시에 브로드캐스트(broadcast)하고 그 데이터베이스들에 의해 제공되는 결과를 사용자에게 웹 문서의 형태로 제시하는 방법을 사용하고 있다. 그러나 이러한 방법은 불필요한 네트워크 자원의 접근과 함께 상당한 통신비용을 발생시킨다. 그러므로 효율적인 분산정보검색을 위해서는 사용자가 원하는 문서를 선택적으로 검색하는 방법이 필요하다. 이러한 방법은 기본적으로 지능적 검색방법을 필요로 하고 있다. 지능적 검색방법은 사용자의 관심도, 선호도를 고려하여 사용자와 관련된 지식을 베이지언 네트워크(Bayesian network) 등으로 표현하고 사용자가 제공하는 피드백 정보를 업데이트하는 형식으로 검색하는 방법, 의미분석을 및 색인어 분류를 통한 자연어 처리 방법, 검색영역의 문서들을 의미적으로 색인하고 검색하는 신경망 메커니즘을 이용하는 방법 등이 있으나 본 연구에서는 SMART, SavvySearch, GLOSS, Retsina, Amalthaea 시스템의 추론방법을 중심으로 살펴보고자

한다[1,3].

SMART 시스템에서는 모든 문서 데이터베이스의 중심 용어 벡터를 포함하는 군집파일(Clustered file)을 색인으로 사용하여 사용자의 질의의 유사도를 측정하는 방법으로 대규모 문서들이 문서 데이터 베이스들에 잘 분류되어 문서 데이터베이스내의 문서들이 균일하게 분포하는 경우에 좋은 검색결과를 가져 올 수 있지만 대부분 용어 벡터공간에서 중심용어들의 분포도를 사전에 예측하기는 어렵기 때문에 종종 잘못된 결과를 가져온다.

GLOSS 시스템에서는 각 용어에 대한 문서빈도나 용어가중치의 합과 같은 통계치를 기반으로 사용자의 질의어와 데이터베이스 내의 각 문서 사이의 관련도를 계산하는 방법을 사용한다. 그러나 이러한 시스템은 각 문서 데이터베이스는 항상 정확한 색인 데이터베이스 정보를 가지고 있어야 한다는 제약적 조건으로 인해 이러한 가정을 만족하지 않는 환경에서는 효과적인 검색 결과를 제공하지 못하는 문제점을 지니고 있다.

SavvySearch 시스템은 사용자의 경험적 정보검색 결과를 이용하여 사용자의 질의어와 각 문서 사이의 관련도를 증강학습법(reinforcement learning method)으로 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.

Retsina 시스템은 인터페이스 에이전트, 작업 에이전트, 정보 에이전트 등으로 에이전트를 구성하고 있다. 인터페이스 에이전트는 사용자와 상호작용을 하며 사용자의 질의를 받아 분석하고 결과를 보여 주는 역할을 수행한다. 작업에이전트는 주어진 작업에 대한 영역지식과 함께 다른 작업 에이전트나 정보 에이전트의 능력을 가지고 사용자가 요구한 작업을 실제 수행하는 에이전트이다. 정보 에이전트는 여러 곳에 흩어진 이형질의 정보소스를 지능적으로 접근할 수 있는 기능을 제공한다.

또한 MIT에서 개발한 Amalthaea 시스템은 멀티 검색엔진으로 구성된 검색 에이전트로, 웹을 탐색하는 웹 검색엔진, 웹에서 특정 정보를 찾아내는 정보검색 에이전트(IRA), 찾아낸 정보를 적절하게 추려내는 정보필터에이전트(IFA)로 구성되며, 오랫동안 사용되지 않은 에이전트는 삭제되고, 자주 사용되는 에이전트들끼리는 서로의 Keyword Vector들을 계승받은 자손 에이전트를 새롭게 생성할 수 있으며 이런 과정에서 IRA와 IFA의 집단은 점차로 특정 사용자의 기호에 맞는 정보를 제공하는 방향으로 학습되어 진다.

표 1은 SMART, GLOSS, Savvy Search 시스템에서 데이터베이스인 DB가 주어졌을 때 주어진 질의 q ∈ Q에 대한 문서 데이터베이스 dbj ∈ DB의 질의관련도 ρdbj(q)를 통해 유사도 평가 함수를 나타내고 있다.

표 1 각 시스템별 관련도 평가함수
Table. 1 Each system degree of association estimation function

Similarity System	질의관련도 ρdbj(q)
SMART	$\frac{\sum_{t_i \in q} \Psi_{db_j}(t_i)}{\sqrt{\lvert q \rvert} \cdot \sqrt{\sum_{t_i \in T} \Psi_{dbj}(t_i)^2}}$
GLOSS	$\frac{\prod_{t_i \in q} \Psi_{dbj}(t_i)}{\lvert D_{dbj} \rvert^{\lvert q \rvert - 1}}$
Savvy Search	$\frac{\sum_{t_i \in q} \Psi dbj(t_i)}{\sqrt{\sum_{t_i \in T} \Psi dbj(t_i)}}$

각 시스템은 이러한 유사도 평가 함수를 통해 지능적 검색기능을 부분적으로 제공하고 있으나 대부분의 정보검색 에이전트에서는 인터넷상의 정보를 처리함에 있어 사용자별 선호도를 고려한 부분에 있어서는 낮은 수준의 인공지능 기술만을 사용하여 기존 검색 엔진과 큰 차별성이 없을 뿐만 아니라 사용자별 적응 학습도 고려되지 않아 개별 사용자의 특성에 맞는 결과를 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 많다[4]. 따라서 본 연구는 사용자의 경험적 지식을 검색에 활용할 수 있는 사례기반 추론기법을 검색 에이전트에 처음으로 적용하고자 한다.

2. 사례기반 추론

사례 기반 추론(Case-Based Reasoning : CBR)은

주어진 문제를 해결하기 위해 과거의 유사한 사례를 바탕으로 문제의 상황에 맞게 응용하여 해를 찾아가는 기법으로 새로운 요구에 대응하는 과거의 해답을 채택하거나, 과거의 사례를 이용하여 새로운 상황을 설명하거나, 과거의 사례로 새로운 해답을 평가하거나, 또는 새로운 상황을 이해하기 위해 선례로부터 주어진 문제에 대한 적당한 해답을 추정하는 작업을 수행한다[5].

CBR은 주어진 문제를 해결하기 위해 필요한 정형적 규칙을 찾기 힘든 문제 영역에 적용하는 것이 유용하며, 특히 과거의 경험으로부터 효과적인 의사결정을 이끌어낼 수 있는 경우에 매우 효과적인 문제해결 방법론이다. 이것은 기억장치에서 현재의 문제와 유사한 해결된 해를 찾고, 과거의 문제와 현재의 문제간의 차이를 고려하여 이전의 해결책들을 현재의 문제에 맞게끔 변형하는 과정을 거친다. CBR을 이용한 방법은 과거의 전문가 시스템에서 사용하던 정형화된 룰(rule)의 추론을 통해서 해를 얻는 방법보다는 단순하면서도 문제 영역이 잘 정형화되지 않는 분야에서는 좋은 접근법이라 할 수 있다. 일반적으로 문제를 해결할 때 미리 모든 지식을 룰로 구축할 수 없는 경우가 많이 있다. 이러한 경우 사례 기반 추론은 주어진 문제가 사례로 저장된 과거의 사실과 같다면 특별한 추론기구의 도움 없이도 해를 도출할 수 있다. 이러한 개념은 문제가 복잡하고 해를 구하는데 많은 시간이 요구되는 문제에서는 과거 사례를 기억하여 찾아 해를 제공해 준다면 해를 얻는 시간이 매우 절약되며 이러한 사례기반 추론을 검색 에이전트에 응용하면 사용자별 특성을 고려한 지능적 시스템을 구축하는데 효율적인 방법론이 될 수 있다[6].

사례기반 추론은 과거 문제에 적용했던 해결책을 수정하여 유사한 새로운 문제의 해결에 사용하므로 문제해결 과정의 재사용을 통해서 자동학습이 가능해 지므로 지능 시스템 구축의 가장 어려운 문제인 지식 습득문제를 자연스럽게 해결할 수 있다. 또한 사례들을 기억장소에 저장해 둔 후 새로운 사례가 들어오면 예전의 사례와 비교하여 기존의 해답을 수정하고 올바른 해답을 찾는 과정을 통해 학습과 지식의 증식을 가능하게 한다.

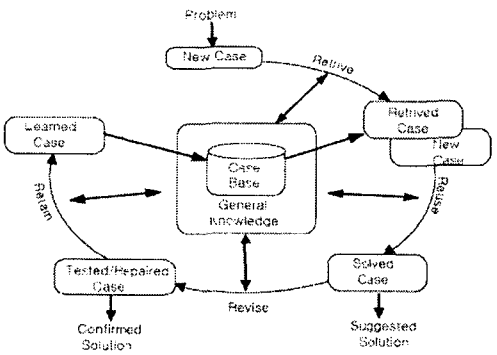

그림 1 개념적인 CBR의 라이프 사이클
Fig. 1 Conceptional CBR's life cycle

사례기반 추론의 개념적 과정은 그림 1과 같이 네 가지 단계의 반복으로 나타난다. 먼저 주어진 문제와 가장 유사한 사례를 찾아오는 회수(retrieve) 단계, 주어진 문제를 과거사례로부터 해결하고자 하는 재사용(reuse)단계, 필요한 경우 유사사례로부터 해를 제시하기 위한 교정(revise)단계, 그리고 교정된 해를 새로운 사례로 저장하는 보유(retain)단계이다. 사례기반시스템의 추론절차는 그림 2에서와 같이 먼저 해결해야할 문제가 주어지면 사례베이스에 저장되어 있는 과거 사례들 가운데 유사한 사례를 조회한다. 조회된 사례가 현재의 상황과 완전히 일치하지 않을 경우 사례 추론기를 통해 이를 수정하여 현재 상황에 맞는 해결책을 제시하는 적응 과정을 거치며 적응과정을 통과한 해결책은 현재 문제에 실제로 적용하는 시험 단계를 거쳐 성공 혹은 실패로 그 결과가 나타난다. 제안된 해결책이 문제해결에 성공한 경우 현재 문제에 대한 데이터를 새로운 사례로 만들어서 사례베이스에 저장하게 된다. 만약 제안된 해결책이 문제해결에 실패하면 실패의 이유를 설명하고 교정규칙을 이용하여 새로운 해결책을 제시한 다음 다시 시험 과정을 거치는 교정 단계가 필요하다.

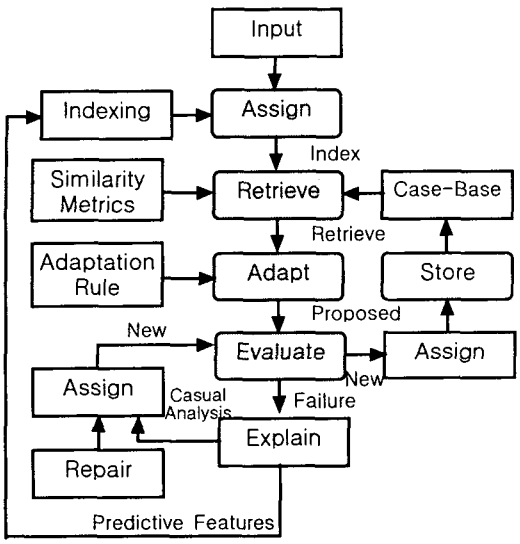

그림 2 CBR을 통한 추론과정
Fig. 2 Reasoning process through CBR

3. 유사도 평가 방법

사례기반 추론에서 사례베이스에 저장된 히스토리 사례와 완전히 일치하는 사례를 찾는 것은 사실 어려우므로 부분적인 일치를 허용하게 되는데 이 부분적인 일치 즉 유사성(similarity)을 어떻게 평가하느냐에 따라서 시스템의 성능이 좌우될 수 있다. 적절한 사례를 평가하는 방법으로 최근접 이웃탐색(the nearest-neighbor search)법이 있다. 이는 새로운 문제의 특성과 사례베이스에 있는 각 사례들과 대응되는 특성을 하나씩 비교하는 매우 간단한 방법이지만 사례베이스의 크기가 증가함에 따라 비용이 급속하게 증가하는 소모적 평가 방법이다. 따라서 본 연구에서는 이 평가함수를 위해 통계적 분석기법의 일종인 유사 군집화(Clustering) 알고리즘을 적용한다. 이 알고리즘은 질의어 q에 대해서 사례베이스 db가 반환하는 관련 문서의 개수를 |db(q)|라 할 때 $|db(q)| = \Pi_{i=1}(q|Ti, Ci, Di)$로 표현하며 이는 질의어 q에 대해 제목 Ti, 카테고리 Ci, 서술항 Di을 갖는 트랜잭션들과 패턴 매칭 작업을 반복적으로 실시할 때 질의어와 일치하는 트랜잭션의 수를 의미한다. 이것은 주어진 예제 질의에 대해서 사례베이스가 관련 문서를 많이 반환하는 경우에는 그 질의를 이루는 각 용어에 대한 사례집합의 유사도가 증가하고, 관련 문서를 반환하지 않는 경우에는 그 질의를 이루는 각 용어에 대한 사례집합의 유사도는 감소한다는 것을 뜻한다. 따라서 충분한 예제질의들에 대해서 이러한 방법으로 각 용어에 대한 사례베이스의 관련도를 계속적으로 조정하여 얻어진 결과를 사용하여 그 질의어와 관련된 정보의 카테고리 집합을 T라하고 유사 질의어 q'가 q'⊆T을 만족하는 개인 및 전체 문서 데이터베이스와의 유사도 SM(q, casei)을 계산하는 평가함수를 다음 식과 같이 정의한다.

$$\text{유사도}(SM) = \alpha * \frac{|db(q)|}{|PH(q)|} + \beta * \frac{|db(q)|}{|AH(q)|}$$

위 식에서 주어진 질의에 대한 사례집합의 유사도는 개인 히스토리 집합의 트랜잭션 수 |PH(q)|와 전체 집합의 트랜잭션 수 |AH(q)|에 반비례하고 관련문서의 수 |db(q)| 및 가중치 α와 β에 비례함을 알 수 있다. 따라서 본 연구에서는 평가함수에 의해 결정된 카테고리 평가값 중에서 최대값을 갖는 카테고리 집합을 그 질의어와 관련성이 가장 높은 유사 카테고리 그룹으로 설정하고 그 카테고리 그룹에 속하는 모든 하부 트랜잭션들을 관련 정보로 제공한다.

Ⅲ. 사례기반 지능형 검색 에이전트의 설계 및 구현

1. 검색 에이전트 시스템의 구성

일반적인 사례기반추론 시스템의 주요 구성 요소는 사례베이스와 문제 해결자이다. 사례베이스는 이전에 해결된 문제의 특성들을 기술한 사례를 저장하며 문제 해결자는 사례검색기(case retriever)와 사례추론기(ca reasoner)로 구성된다. 어떤 문제가 주어지면 사례검색기는 사례베이스에서 가장 적절한 사례를 식별하고 유사도를 평가하기 위해 사례추론기를 이용한다. 일반적으로 이미 조회된 사례를 조사하고 필요한 경우 적응과정을 거쳐서 새로운 문제의 해결을 시도한다. 그러나 응용분야의 특성이나 사례베이스의 내용에 따라서 사례추론기가 필요치 않을 수도 있다. 즉 거의 정확한 사례를 찾는 경우 이를 사용자에게 제시하고 사용자는 필요한 약간의 수정을

가하도록 할 수도 있다.

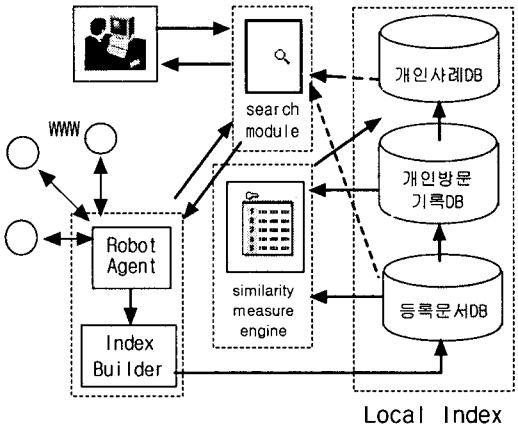

그림 3 제안 시스템의 구조
Fig. 3 Structure of proposal system

그런데 본 연구에서 구현한 사례기반 검색 에이전트는 그림 3과 같이 사용자의 검색요구에 따라 관련된 정보를 찾아 보여 주는 검색 모듈 부분(search module), 유형별 사례 정보를 저장하고 있는 색인 데이터베이스 부분(local index), 축적된 사례를 바탕으로 주어진 검색어의 부합 정도를 측정하여 검색어가 속하는 카테고리를 결정하는 유사도 측정 엔진부분(similarity measure engine)과 웹으로부터 관련 문서를 가져오는 로봇에이전트 부분(robot agent)으로 이루어져 있다.

검색 모듈 부분은 다시 프리젠테이션 계층(presentation)과 트리거(trigger) 계층으로 구성된다. 프리젠테이션 계층은 사용자의 질의어와 색인 구축기를 통해 구성된 색인 데이터베이스의 주제어를 비교하여 일치하면 해당 정보를 HTML 문서 형태로 표현하여 사용자에게 보여준다. 그리고 트리거 계층은 로봇 에이전트와 연동하여 상호구동적(interactive)으로 동작하는 계층이다. 트리거 계층은 사용자의 질의 검색어가 색인 데이터베이스의 주제어로 존재하지 않을 경우 로봇 에이전트에게 URL 데이터베이스의 등록 도메인 이름을 이용하여 웹에서 다시 관련된 내용을 검색하여 색인 데이터베이스를 재구성하도록 요구한다. 색인 데이터베이스 부분에서는 로봇에이전트로부터 인터넷상의 사이트 정보들을 추출하여 등록문서 데이터베이스가 구성되며 개인방문기록 데이터베이스는 개인별 검색 패턴을 기억하며, 개인사례 데이터베이스에는 개인방문기록 데이터베이스로부터 유사도 측정 알고리즘에 의해 유추된 카테고리 그룹정보를 기억하는 테이블 집합이다. 유사도 측정 엔진은 유사도를 계산하는 유사 군집화 알고리즘에 따라 개인방문기록을 바탕으로 개인별 카테고리 그룹을 결정하고 이 결과를 개인사례 데이터베이스에 저장하는 기능을 담당한다.

로봇 에이전트는 URL 데이터베이스로부터의 사이트 정보를 참조하여 해당 웹 서버들을 탐색하여 원시자료(raw Data)를 수집하고 이 자료는 색인 구축기에 전달되어 등록문서 데이터베이스 구축과 지역데이터베이스(local index)에 존재하지 않는 정보 즉 트리거된 정보를 웹으로 가져오는데 이용된다.

2. 검색어 처리 절차

검색어 처리과정은 그림 4에서 나타낸 것과 같이 다음과 같은 단계를 거쳐 처리된다.

- 단계 1: 먼저 사용자로부터 질의어가 입력된다.
- 단계 2: 주어진 질의어와 일치하는 카테고리 그룹을 개인사례정보 테이블에서 검색한다.
- 단계 3: 일치하는 사례가 발견되면 선정된 유사 카테고리 군집에 속하는 모든 하부 트랜잭션들을 사례기반 검색정보로서 제공하고 일치하는 카테고리 집합이 존재하지 않으면 등록문서 테이블에서 전체 사용자의 정보를 바탕으로 유사 카테고리 그룹을 계산하여 그 군집에 속하는 하부 트랜잭션을 사례기반 검색 정보로 제공한다. 만약 등록문서 테이블에서도 관련 정보를 발견하지 못하면 검색에이전트는 로봇에이전트를 트리거 시켜 웹으로부터 단순 패턴 비교하여 관련정보를 사용자에게 보여준다. 이 정보는 지능적 추론 기능이 배제된 검색 결과이므로 검색어와의 유사도 정도에 대한 신뢰도를 보장받을 수 없다.
- 단계 4: 새롭게 방문한 문서는 사례베이스를 변경시키게 되어 개인방문 테이블에 사용자의 방문 정보가 추가된다. 또한 웹으로부터 추가된 사례는 등록문서 테이블에 축적된다. 따라서 변경된 사례 정보를 재사용하여 카테고리 그룹을 다시 작성하도록 한다..
- 단계 5: 추가된 사례를 바탕으로 카테고리 그룹을 다시 결정하기 위해 유사 군집화 알고리즘을 이용하여 유사도를 계산하여 결정된 카테고리 정보를 새로운 카테고리그룹으로 선정한다.

■ 단계 6: 새로운 카테고리 그룹 정보와 기존의 카테고리 그룹 정보를 비교하여 정보에 차이가 있으면 카테고리 그룹정보를 개선하고 새롭게 결정된 카테고리 그룹을 새로운 사례정보로 개인사례정보 테이블에 보유한다.

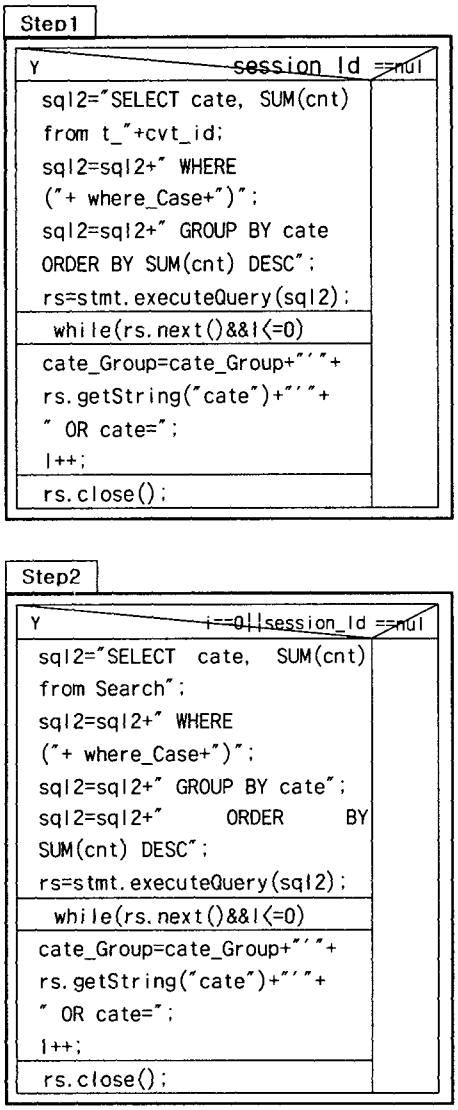

그림 4 사례기반 검색알고리즘
Fig. 4 Case based search Algorithm

Ⅳ. 실험 및 검색성능 평가

1. 실험결과

본 연구에서는 사례기반 추론을 이용한 지능적 검색 정도를 실험하기 위해 검색 엔진을 구현하고 주어진 질의어와 관련 문서 사이에 유사도 정도를 평가하였다. 그림 5는 "모터"라는 질의어에 따른 사례기반검색 결과로 총 19개의 관련 문서를 출력하는 화면이다. 이 그림에서 사용자가 선택할 수 있는 메뉴를 사례기반검색에서부터 우측으로 AI-SEA일반검색, 전문웹검색 및 웹페이지 검색 순으로 구성시켜 사례기반 검색뿐만 아니라 일반검색 및 범용검색에 대한 처리도 고려하였다.

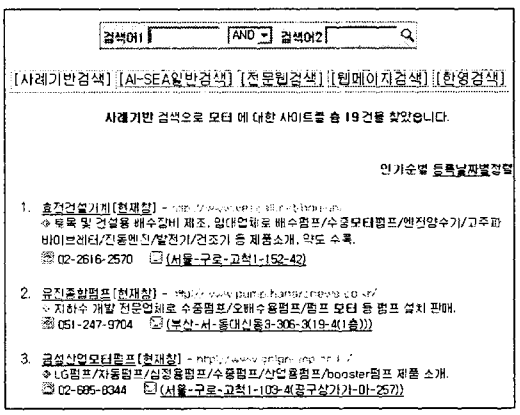

그림 5 검색에이전트에서 검색어 처리 결과
Fig. 5 Search word processing result in search a

2. 실험환경

사례기반의 지능적 검색에 대한 실험을 위해 해양관련 정보를 저장하고 있는 로컬 색인 데이터베이스로부터 수집한 7760개의 문서 자료를 이용하였다. 이 문서들은 전체 12개의 대분류 카테고리로 분류되어 있고 대분류 카테고리 밑에 최고 7개의 하부 카테고리 그룹 레벨을 가지고 있다. 실험에 사용된 검색어 개수는 총 30개이며 검색 카테고리 그룹의 개수는 1911개이다.

정보검색에서 성능평가는 일반적으로 검색된 결과의

정확률(Precision)과 재현률(Recall)의 측정으로 이루어진다. 정보검색 성능 평가의 척도로서 정확률과 재현률을 다음과 같이 정의하였다.

$$\text{정확률} = \frac{\text{주어진 질의어와 관련된 검색 문서 개수}}{\text{검색된 전체 문서들 개수}}$$

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

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 정확률은 주어진 검색어의 종류에 따라 다른 결과를 나타낼 수 도 있다. 따라서 두 성능 평가율을 일정한 값으로 수렴시키기 위해서는 여러 카테고리 집합에서 균일한 분산을 갖는 실험 데이터의 개수를 증가시킬 필요가 있고, 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.

3. 성능분석

본 연구에서의 실험은 네 가지 검색방법으로 수행된다. 첫 번째는 사례기반검색으로 사용자별 개인사례정보를 바탕으로 유사도가 가장 높은 카테고리 그룹의 정보만을 제공하는 경우이다. 두 번째로 AI-SEA일반검색은 해양관련내용으로 필터(filter)된 색인DB 정보를 대상으로 검색을 수행하는 과정이며, 세 번째는 전문웹검색으로 상용검색엔진에서 카테고리별로 분류된 정보를 대상으로 검색을 수행한 경우이다. 네 번째는 웹페이지 검색으로 범용검색엔진으로 일반 웹 문서전체를 대상으로 검색을 수행하는 경우이다. 실험은 네 가지 검색방법에 대해 해양관련 주제로 30개의 질의를 사용하여 검색을 수행하고 나온 결과를 관련수준(Relevant level)별로 나눈 5개씩의 질의 결과에 대한 평균치를 정확률과 재현률로 표 2에 나타내었다. 전문 웹검색과 웹페이지 검색에서 재현률을 나타낼 수 없는 이유는 검색에 이용된 범용검색엔진에서 보유하는 데이터베이스의 관련 문서 개수를 산출할 수 없기 때문이다.

표 2 검색방법에따른 정확률과 재현률
Table. 2 The correctness rate and the recall rate f search method

Level	사례기반 검색		AI-SEA 일반검색		전문웹검색	웹페이지 검색
	정확률	재현률	정확률	재현률	정확률	정확률
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

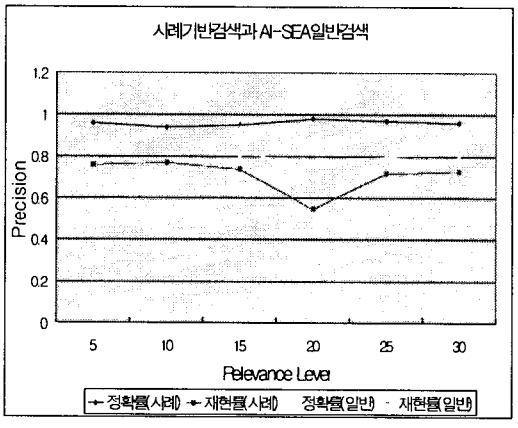

그림 6 사례기반검색과 일반검색의 정확률과 재현률의 변화추이
Fig 6 The change trend of the correctness rate and the r case based search and the general search

그림 6은 표 2의 자료 중 사례기반검색과 AI-SEA일반 검색에서 정확률과 재현률의 변화추이를 보여 주고 있다. 그래프 결과에 따르면 일반적으로 정확률과 재현률은 상대적으로 역비례 관계를 가지는 것으로 나타났다. 또한 사례기반검색이 AI-SEA검색보다 정확률이 높고 재현률은 AI-SEA일반검색이 사례기반검색에 비해 평균적으로 높은 것으로 나타났다. 이것은 개인의 검색성향을 고려한 사례기반검색이 한정된 카테고리 그룹 정보를 집중적으로 추출하기 때문인 것으로 보인다.

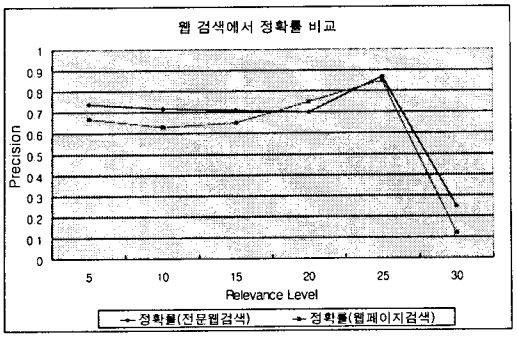

그림 7 전문웹검색과 웹페이지 검색에서의 정확률의 변화 추이
Fig. 7 The change trend of the correctness rate in s search and web page search

그림 7은 전문웹검색과 웹페이지 검색의 정확률을 비교하고 있다. 평균적으로 전문웹검색이 일반 문서전체를 대상으로 하는 웹페이지 검색에 비해 상대적으로 높은 정확률을 나타내고 있다. 이것은 전문웹검색이 주제별로 분류된 영역에서 검색을 수행하기 때문이다. 이 그림에서 정확률이 큰 폭으로 감소하고 있는 경우가 있는데 이것은 특정한 주제어그룹에서 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 정확률을 현저히 낮출 가능성이 있다.

V. 결론

본 논문은 웹 검색에이전트에서 사용자의 과거 정보검색 사례를 기반으로 유사도를 평가하여 사용자의 검색 의도를 충족하는 지능적 검색 방법을 제안하였다. 유사도 측정을 위해서는 통계적 분석기법의 일종인 유사 군집화(Clustering) 알고리즘을 사용하여 평가함수에 의해 결정된 카테고리 평가값 중에서 최대값을 갖는 카테고리 집합을 그 질의어와 관련성이 가장 높은 유사 카테고리 군집으로 설정하고 그 카테고리 군집에 속하는 모든 하부 트랜잭션들을 사례기반 검색 정보들로 제공하여 사용자의 의도에 맞는 정보 검색 능력을 높였다.

지식의 확장과 카테고리 그룹의 갱신은 사례베이스 관리 모듈이 유사사례추출과 적응단계에서 여과된 정보를 사례표현 단계로 적절히 피드백 시켜 사용자의 사례그룹을 지속적으로 변경시키는 학습을 통해 이루었다.

제안된 사례기반 검색에이전트를 이용하여 실시한 정확률과 재현률에 대한 평가 실험에서는 평균적으로 정확률과 재현률은 역비례 관계를 가지는 것으로 나타났으며 사례기반검색과 AI-SEA일반 검색에서의 정확률과 재현률의 변화추이 비교에서는 전체적으로 볼 때 사례기반검색이 AI-SEA검색보다 정확률이 높고 재현률은 AI-SEA 일반검색이 사례기반검색에 비해 평균적으로 높은 것으로 나타났다. 이것은 각 개인의 검색성향을 통계적으로 고려한 사례기반 검색이 정확률에서 우수한 결과를 가져 올 수 있음을 입증한 것으로 사료된다.

참고문헌

[1] 최용석, "분산된 웹 데이터베이스에서의 정보검색 신경망 에이전트", 박사학위논문, 2000.

[2] 김은경, "규칙기반추론과 사례기반추론의 통합에 관한 연구", 한국교육기술대학원 논문집 제4권 제1호, 1995.

[3] Maarek Y., "Berry D. and Kaiser G, An Information Retrieval Approach For Automatically Construction Software Libraries", IEEE Transaction On Software Engineering, Vol. 17, No. 8, pp.800-813, August 1991.

[4] R. Armstrong et.al., "webWatcher:Machine Learning and Hypertext", Fachgruppentreffen Maschinelles Lernen, Dortmund, Germanu, 1995.

[5] Aamodt, A. & Plaza, E., "Case-Based Reasoning: Foundation al Issues, Methodological variations, and System Approach, AI Communications, Vol.7, No.1, 1994.

[6] Marko Balabanovic and Yoav Shoham, "Learning Information Retrieval Agents: Experiments with Automated Web Browsing" in Proceedings of the AAAI Spring symposium on Information Gathering from Heterogeneous, Distributed Resources, March 1995.

저 자 소 개

하 창 승

1984년 2월 한국해양대학교 항해학과 졸업 (공학사)

1992년 2월 한국해양대학교 전자통신공학과 (공학석사)

2001년 2월 한국해양대학교 전자통신공학과 (공학박사 수료)

1996년 9월 - 현재 동명대학 정보통신계열 조교수

류 길 수

1976 2월 한국해양대학교 기관학과 졸업 (공학사)

1979 2월 한국해양대학교 대학원 기관학과 (공학석사)

1986 일본동경공업대학 대학원 (공학석사-정보공학)

1989 일본동경공업대학 대학원 (공학박사-정보공학)

1982 - 현재 한국해양대학교 기계정보공학부 교수