

계층적 웹 환경에서의 멀티-에이전트 기반 웹 마이닝 시스템 설계

Modeling a Multi-Agent based Web Mining System on the Hierarchical Web Environment

윤희병* · 김화수**

Heebyung Yoon and Hwa-Soo Kim

* 국방대학교 전산정보학과

** 아주대학교 정보통신대학원

요 약

웹 기반하에서 사용자의 질의에 대한 효율적인 검색결과를 제공하기 위하여 다양한 검색 알고리즘들이 개발되어 왔으며, 이러한 알고리즘들의 대부분은 사용자의 선호도나 편의성을 고려하였다. 그러나 지금까지 개발된 검색 알고리즘들은 일반적으로 웹이라는 수평의 비계층적인 웹 환경에서 개발된 것으로서 기업의 전사적 네트워크와 같이 계층적이고 기능적으로 복잡하게 구성되어 있는 웹 기반 환경에서는 적용하기가 힘든 실정이다. 본 논문에서는 이러한 특수한 웹 기반 환경하에서 사용자에게 효율적으로 마이닝 결과를 제공할 수 있는 멀티 에이전트 기반의 웹 마이닝 시스템을 제안한다. 이를 위해 우리는 계층적 웹 기반 환경이라는 네트워크 모델을 제시하며, 제시된 웹 환경에서 적용할 수 있는 4개의 협력 에이전트와 14개의 프로세스 모듈을 가진 멀티 에이전트 기반의 웹 마이닝 시스템을 설계한다. 그리고 각 에이전트에 대한 세부기능을 계층적 환경을 고려하여 모듈별로 설명하며 특히, 새로운 머징 에이전트와 개선된 랭킹 알고리즘을 그래프 이론을 적용하여 제안한다.

Abstract

In order to provide efficient retrieving results for user query on the web environment, the various searching algorithms have developed and considered user's preference and convenience. However, the searching algorithms are developed on the horizontal and non hierarchical web environment in general and could not apply to the complex hierarchical and functional web environments such like the enterprise network. In this paper, we purpose the multi-agent based web mining system which can provide the efficient mining results to the user on the special web environment. For doing this, we suggest the network model with the hierarchical web environment and model the multi agent based web mining system which has four corporation agents and fourteen process modules. Then, we explain the detailed functions of each agent considered the hierarchical environment according to the module. Especially, we purpose the new merging agent and improved ranking algorithm by using the graph theory.

Key words : 멀티 에이전트, 웹 마이닝 시스템, 계층적 웹 환경, 랭킹 알고리즘, 머징 에이전트.

1. 서 론

최근에 웹상에서 사용자가 원하는 신뢰성이 있는 정보를 효율적으로 찾아 주는 웹 마이닝 시스템에 대한 관심이 증가하고 있다. 그러나 웹상의 대량의 정보 속에서 사용자 스스로 원하는 주제에 대한 최선의 정보를 좀 더 쉽고 정확하고 신속하게 얻는다는 것은 매우 어렵다. 따라서 이러한 사용자의 요구를 대행하여 효율적으로 작업을 수행해주는 에이전트라는 분야에 대한 관심이 급증하고 있다.

특히 웹 마이닝 분야는 하나의 에이전트로 해결하지 못하는 복잡한 기능이 필요하므로 여러 에이전트간의 협동이 필

요하게 되었고 이를 효과적으로 수행하기 위하여 멀티 에이전트라는 개념이 나오게 되었다. 이러한 멀티 에이전트 시스템에 대한 연구가 여러 분야에서 활발하게 진행되고 있으며, 이에 대한 응용 분야를 살펴보면 지능형 홈 네트워크 서비스, 지능형 빌딩 감식과 통제, 전자상거래, 비즈니스 절차 관리, 그리고 정보검색 등이 있다.

그러나 이들 연구의 대부분은 우리가 흔히 접하는 인터넷이라는 공개된 웹 기반의 환경, 또는 하나의 지역이나 빌딩 등으로 제한되어 있는 웹 환경을 기반으로 하고 있다. 따라서 지역 및 기능적으로 분산이 되어 있고 또한 계층적으로 구성이 되어 있는 기업의 전사적 네트워크와 같은 특수한 웹 환경에서 적용된 사례는 웹 마이닝 분야에서는 찾아보기가 어렵다.

계층적 웹 환경에서의 효율적인 웹 마이닝 시스템을 설계하기 위해서는 웹 사이트마다 시스템을 각각 운영하는 것보

접수일자 : 2003년 8월 27일
완료일자 : 2003년 11월 18일

다 주어진 웹 환경에 맞는 시스템을 설계하여 운용하는 것이 더 효율적이다. 또한 웹 마이닝 시스템의 모든 기능을 분석하여 이를 독립적으로 운용이 가능한 에이전트로 모듈화하고 이를 통합한 멀티 에이전트 기반의 웹 마이닝 시스템을 설계하여야 한다. 이렇게 함으로써 사용자의 질의에 가장 효과적으로 응답할 수 있는 시스템 설계가 가능해진다.

따라서 본 논문에서는 전사적 네트워크와 같이 계층적이고 기능적으로 복잡하게 구성되어 있는 특수한 웹 기반 환경에 가장 적합한 멀티-에이전트 기반의 웹 마이닝 시스템을 설계하기 위하여 제 2장에서는 멀티-에이전트의 구조, 검색 엔진, 그리고 랭킹 알고리즘에 대하여 살펴본다. 제 3장에서는 계층적 웹 환경 및 웹 마이닝 시스템을 설계하기 위하여 고려해야 될 사항에 대하여 계층적인 웹 환경 측면과 정보검색 시스템의 기능 측면 등 두 가지 측면에 살펴본다. 제 4장은 제시된 여러 고려사항 및 특수한 환경을 고려하여 효율적인 웹 마이닝 시스템을 설계하기 위한 시스템 구조와 기능을 제안하며 제 5장은 이를 이용한 구현방안을 제시하고, 결론을 맺는다.

2. 관련 연구

2.1 멀티-에이전트 구조

멀티 에이전트란 하나의 에이전트로 해결하지 못하는 복잡한 문제를 여러 개의 에이전트간의 협력작업을 통해 해결하는 에이전트를 말한다. 여러 에이전트간 협력작업을 위한 에이전트의 내부구조, 연결구조 그리고 정보전달 방식 등을 멀티 에이전트의 3대 구성요소라 한다.

멀티 에이전트 구조의 대표적인 사례가 QMW 대학의 ARCHON[1]이다. 이것은 모든 에이전트가 서로 다른 에이전트에 대한 정보를 보유하여 자신이 원하는 서비스를 직접 에이전트에 요청하는 구조이다. 이에 반해 EMAF[2]는 조정자 역할을 하는 조정 에이전트를 통해 서비스를 주고받는 구조이다. 예를 들면, 조정 에이전트는 요청한 서비스 내용을 분석해 이를 수행할 에이전트를 선정하고 해당 에이전트에 요구된 서비스 메시지를 전달한다. 이때 서비스를 요청한 에이전트나 서비스를 제공한 에이전트는 상대방이 누구인지 모르고 동작하게 된다.

그리고 ARCHON과 EMAF를 혼합한 구조도 있다. 이 구조는 조정 에이전트가 하부 에이전트의 정보를 갖는 것은 EMAF와 같다. 그러나 조정 에이전트가 직접 서비스를 수행하지 않고 하부 에이전트간 연결 채널을 만들어 준다는 것이 다르다. 따라서 하부 에이전트들은 각 에이전트들에 대한 정보를 가질 필요가 없다는 것이 ARCHON과 다르다.

1996년에 에이전트 관련 최초의 세계적인 표준화 기구인 FIPA[3]에서 제안한 AP(Agent Platform)가 있다. 이 AP는 내부 에이전트간의 메시지 전송 및 통신을 지원하는 ACC(Agent Communication Channel), 에이전트의 이름과 주소 정보를 에이전트에게 제공하는 ANS(Agent Name Server), 내부 에이전트 능력이나 서비스 정보를 타 에이전트에게 제공하는 DF(Directory Facilitator), 그리고 내부 에이전트에 대한 등록, 제거, 일시정지, 회복 등의 관리 기능을 수행하는 AMS(Agent Management System) 등으로 구성되어 있다.

2.2 검색 엔진

검색엔진은 크게 사용자 입장에서의 분류와 세대별 분류로 구분할 수 있다. 사용자 입장에서는 주제별 검색엔진, 단

어별 검색엔진, 메타 검색엔진으로 분류된다. 주제별 검색엔진은 초보자들이 가장 많이 사용하는 방법으로 주제별로 각 페이지들이 잘 정리된 검색엔진이다. 이것은 정보에 대한 지식이 분명치 않거나 어떤 분야에 대한 검색을 원할 경우에 사용된다. 단어별 검색엔진은 하나의 DB에 모든 URL을 저장하고 특정 키워드를 입력하여 원하는 정보를 찾는 방법이다. 요즘 대부분의 검색엔진은 주제별 검색엔진과 단어별 검색엔진을 모두 제공한다. 메타 검색엔진이란 한번의 검색으로 여러 검색엔진을 동시에 검색하는 방법으로 대부분의 멀티-스레드 기법을 사용한다. 이것은 가장 짧은 시간에 원하는 정보를 얻기 위하여 여러 개의 검색엔진을 통합하는 방법이다[4].

세대별 분류는 세대 전 검색엔진, 1세대 검색엔진, 2세대 검색엔진, 3세대 검색엔진으로 분류된다[5]. 세대 전 검색엔진은 인터넷 검색 이전의 시기로 주로 어떤 일정한 기관 내의 자료를 검색하는 DB 검색용이다. 1세대 검색엔진은 웹 로봇이 인터넷을 돌아다니며 자동으로 웹 문서를 수집해오는 엔진으로 자동 색인을 위해서 형태소 분석을 사용하나 초보적인 수준이다. 2세대 검색엔진은 사용자의 취향을 검색에 반영하거나 형용사나 동사도 처리하고 또한 멀티 워드로 된 질의어도 처리할 수 있도록 기술적으로 발전된 세대이다. 3세대 검색엔진은 현재 개발 중인 엔진으로 구문 분석을 할 수 있다는 것이 2세대와 큰 차이점이다[6].

2.3 랭킹 알고리즘

웹 마이닝 시스템에서 문서에 점수를 부여한다는 것은 사용자의 선호도나 편의성을 고려한다는 의미와 같다. 이러한 랭킹 알고리즘은 거의 대부분 문서의 내용이나 또는 하이퍼링크 구조를 분석하는 방법으로 개발되어 왔다. 이들 중 가장 대표적인 방법이 텍스트 기반의 Vector-Space 모델[7]과 하이퍼링크 구조 기반의 HITS[8]와 PageRank[9]가 있다.

여기서 하이퍼링크를 다시 질의-종속과 질의-독립의 두 가지 범주로 분류할 수 있다[10]. 질의-종속에는 HITS, 질의 용어를 앵커 텍스트와 비교하여 가중치가 포함된 개정 규칙을 얻기 위한 HITS 확장[11,12], 그리고 링크 분석을 위한 확률론적 접근방법으로서 두개의 매트릭스를 이용하는 SALSA[13]가 있다. 질의-독립에는 PageRank, 웹 페이지 점수 시스템으로서 특정 주제에 중점을 두고 페이지의 상대적인 랭킹을 계산하는데 목적이 있는 Focused PageRank[14] 등이 있다. 여기서 PageRank 알고리즘은 실제 웹 사이트의 크기를 고려하여 사이트 전체를 랭크하지 않고 각 페이지에 대하여 랭크를 결정하는데 대략적이고 반복적인 계산을 수행한다. 페이지 A에 대한 PageRank를 구하는 식은 다음과 같다[9].

$$PR(A) = (1 - d) + d \left(\frac{PR(T_1)}{C(T_1)} + \dots + \frac{PR(T_n)}{C(T_n)} \right) \quad (1)$$

여기서 PR(T_i)는 페이지 A에 연결되는 페이지 T_i의 PageRank이며, C(T_i)는 페이지 T_i에서 외부로 연결되는 링크의 수를 말하며, d는 0과 1사이의 값을 갖는 댄핑(damping) 요소를 말한다.

3. 계층적 웹 환경 및 설계 고려사항

계층적 웹 환경이란 기업의 전사적 네트워크를 모델로 하여 고려한 환경이다. 중앙에 하나의 중심노드가 있고 중심노드 밑에 지역적으로 핵심기능을 수행하는 여러 개의 지역핵

심노드, 하위에 기능적으로 핵심기능을 수행하는 기능핵심노드, 그리고 각각의 독립된 기능 즉, 교육, 관리, 인사 등을 수행하는 많은 수의 가지노드가 계층적, 기능적으로 구성되어 있는 트리형태의 논리적인 웹 구조를 말한다. 이러한 특수한 형태의 웹 구성도가 그림 1에 도시되어 있다.

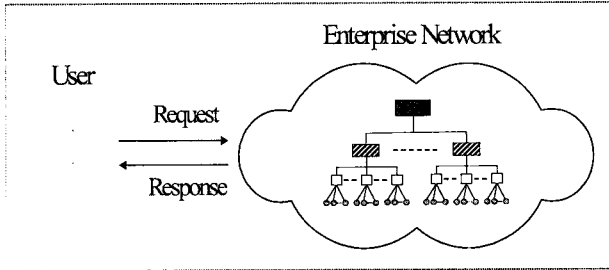


그림 1. 계층적 웹 환경.
Fig. 1. Hierarchical web environment.

따라서 이러한 특수한 환경에 속해 있는 사용자들은 자신의 현 위치에 대한 정보들을 검색에 반영시켜 좀 더 쉽고 정확하고 신속하게 검색결과를 얻기 원할 것이다. 사용자 관점에서 효율적인 정보검색 시스템을 설계하기 위한 고려요소들 계층적인 웹 환경 측면과 웹 마이닝 시스템의 기능 측면 등 두 가지 측면에 살펴볼 수 있다.

먼저 계층적인 웹 환경 측면에서는 각각의 독립적인 기능들을 수행하는 에이전트들을 어디에 배치할 것인가와 각 에이전트의 기능 수행 도중에 구축된 인덱스 DB를 어느 노드에 어느 수준까지로 배치할 것인가가 주요한 고려요소가 된다. 그리고 검색로봇이 네트워크에 연결되어 있는 수많은 웹 사이트를 주기적으로 돌아다니며 웹 문서를 수집하는 크롤링(crawling) 기법도 계층적인 웹 환경에서는 고려해야 될 것이다.

웹 마이닝의 기능 측면에서는 에이전트의 분류, 에이전트의 기능 식별, 에이전트간 협력, 그리고 검색된 문서에 가중치를 부여하는 랭킹 알고리즘 등이 주요한 고려요소가 된다. 이러한 고려요소들 살펴보면 웹 서버의 QoS를 보장해주기 위한 효율적인 URL 할당방법, 웹 문서의 크롤링 주기를 짧게 해주는 재생패치 전략 및 재생주기를 위한 타임스탬프, 동일한 문서의 중복저장을 방지하는 콘텐츠 확인시험(content seen test), 기능적 핵심노드간 검색로봇의 중복접근 방지전략, 검색로봇이 수집한 문서에 대한 식별번호 부여, 문서의 랭킹 알고리즘을 위한 노드별 가중치 값 부여 등이 있으며 그 외에 사용자의 계층별 기능별 노드의 위치를 고려한 IP 주소식별 및 식별된 IP 주소를 이용한 문서의 재정렬 기능, 그리고 인덱스 에이전트의 갱신시간 접수 및 하위 웹 사이트의 인덱스 DB 머지 등이 주요한 고려요소가 될 것이다.

4. 웹 마이닝 시스템 설계

4.1 시스템 구조

주어진 계층적 웹 기반 환경에서 사용자의 질의에 가장 효율적으로 응답할 수 있는 웹 마이닝 시스템을 모델링하기 위해 우리는 다음의 네 가지 에이전트로 분류한다. 사용자 인터페이스 에이전트(UIA), 머지 에이전트(MA), 인덱스 에이전트(IA), 그리고 로봇 에이전트(RA)이다.

UIA는 사용자의 요청 정보를 분석하고 검색결과를 사용

자에게 제공한다. MA는 하위 웹 사이트의 인덱스 DB를 머지하고 사용자의 조건에 따라 인덱스 단어를 제공한다. IA는 검색된 문서를 인덱스하여 DB에 저장하고 문서를 랭킹한다. 그리고 RA는 URL을 할당받아 웹 사이트로부터 문서를 가져온다. 이와 관련한 전반적인 멀티-에이전트 기반의 웹 마이닝 시스템 구성도가 그림 2에 도시되어 있다. 참고로 이 그림에서는 웹 마이닝 시스템에 대한 에이전트 기능을 중심으로 하여 분류하였으며 에이전트간 정보교환을 위한 통신프로토콜 및 구조는 생략하였다.

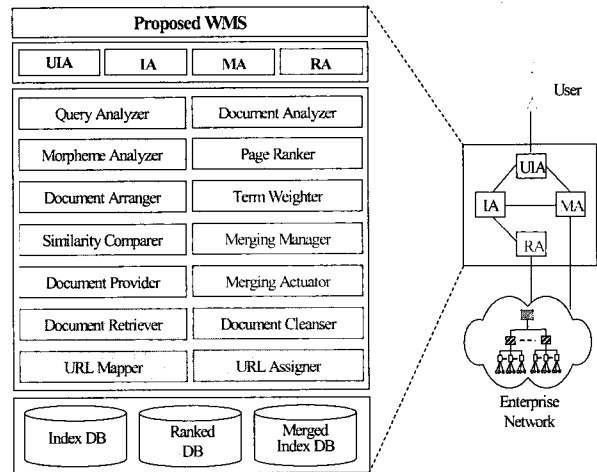


그림 2. 제안한 멀티-에이전트 기반의 웹 마이닝 시스템 구성도.

Fig. 2. Purposed multi-agent based web mining system configuration.

그림 2에서 제안한 웹 마이닝 시스템은 세 개의 계층으로 구성되어 있다. 제일 상위 계층은 웹 마이닝 시스템 계층이며 그 아래가 에이전트계층이고 다음이 프로세스 모듈 계층이다. 에이전트 계층에는 네 개의 에이전트와 세 개의 DB가, 프로세스 모듈 계층에는 14개의 프로세스 모듈이 각각 구성되어 있다. 각 모듈은 서로 관련된 기능으로 그룹화가 가능하며 이들 그룹이 서로 정보를 교환하도록 협력하는 구조를 갖게 되는데 이것이 바로 멀티-에이전트 기반의 웹 마이닝 시스템이다.

4.2 시스템 기능

UIA는 다섯 개의 모듈로 구분할 수 있다. Query Analyzer는 사용자의 질의를 받아 분석하며 형태소 분석을 위하여 질의 단어를 Morpheme Analyzer로 제공한다. 또한 사용자의 현 IP 주소를 Document Arranger에 제공한다. Morpheme Analyzer는 사용자의 질의 단어를 분석하여 여러 문서에 동일하게 많이 사용되는 단어와 적게 사용되는 단어를 제거한다. 즉, in, the, a, with 등과 같은 단어는 색인어로서 의미가 없으며 사용자의 오타 등은 적게 사용되는 단어로서 판단하여 불용어(stop word)로 처리한다. 불용어가 제거되면 다음은 동일한 의미의 자연어로 일관된 형태의 색인어로 만들기 위해 어간 정제과정을 거쳐 합성어를 제공한다. Similarity Comparer는 사용자의 질의어가 합성어로 입력되면 IA의 Index DB, MA의 Index DB의 문서와 유사도를 비교하여 검색된 문서의 식별번호를 제공한다. Document Arranger는 문서의 ID로 우선순위가 부여된 문서를 찾아내며 사용자의 IP 주소를 고려하여 최종적인 문서를 우선순위

별로 제공한다. Document Provider는 사용자에게 최종적으로 순위화된 문서를 페이지 당 10개 또는 20개 등으로 구분하여 URL, 콘텐츠, 유사도 등과 함께 제공한다. 이와 같은 사용자 인터페이스 에이전트의 모듈에 대한 기능 구성도가 그림 3에 도시되어 있다.

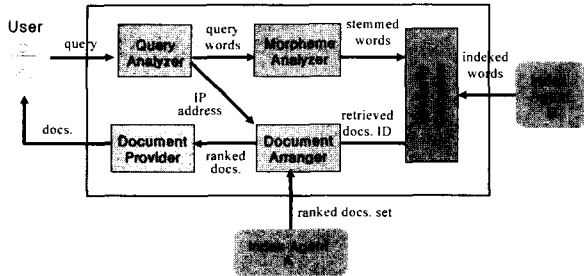


그림 3. UIA의 기능 구성도.
Fig. 3. Functional Configuration of UIA.

MA의 Merge Manager는 사용자의 요청 정보를 받아 분석하며 인덱스 단어를 제공한다. Merge Actuator는 IA로부터 갱신 신호를 받아 관련이 있는 하위 웹 사이트의 인덱스 DB를 머지하며 UIA에 인덱스 단어를 제공한다.

IA의 Document Analyzer는 수집된 웹 문서의 태그 및 내용을 분석하고 하이퍼링크 구조와 합성어를 추출하여 Page Ranker와 Term Weighter로 각각 보낸다. Term Weighter는 Vector-Space 모델[7]의 $tf \times idf$ 를 이용하여 색인어의 가중치를 계산하며 문서의 ID 순서로 가중치와 인덱스 단어를 포함하여 인덱스 DB에 저장한다. 여기서 tf (term frequency)는 문서내에서의 색인어의 빈도를 나타내며 idf (inverse document frequency)는 문서에 사용된 색인어가 다른 문서에 많이 등장하는 단어일수록 그 가중치가 줄어들 수 있도록 전체 문서에 대한 문서내의 색인어를 표현한 것이다. Page Ranker는 문서의 하이퍼링크 구조와 문서가 수집된 출처에 대한 노드의 위치정보를 이용하여 각 문서에 점수를 부여한다. 이와 같은 인덱스 에이전트의 기능 구성도가 그림 4에 도시되어 있다.

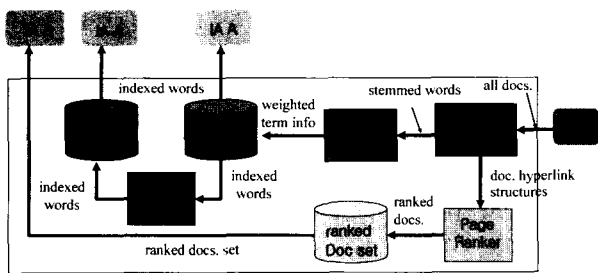


그림 4. IA의 기능 구성도.
Fig. 4. Functional Configuration of IA.

RA의 Document Retriever는 검색로봇에게 제공할 URL 주소를 받아서 DNS로부터 IP 주소로 변환하여 로봇에 제공한다. Document Cleanser는 로봇이 수집한 문서의 하이퍼링크 구조를 분석하여 URL 정보, 이미지, 태그수 등의 정보를 추출한다. 여기서 새로운 URL이 발견된 경우에는 URL 큐에 삽입한다. URL Mapper는 로봇에게 정확한 문서위치를 제공하기 위하여 문서의 상대주소를 절대주소로 변환한다. URL Assigner는 웹 서버의 QoS를 보장하기 위하여 적합한

URL 주소를 로봇에게 할당하는 것으로 검색로봇이 주기적으로 웹 사이트를 탐색하여 문서를 수집하기 위해서는 계층적 웹 구조를 고려한 할당 알고리즘이 개선되어야 한다. 이와 같은 로봇 에이전트의 모듈에 대한 기능 구성도가 그림 5에 도시되어 있다.

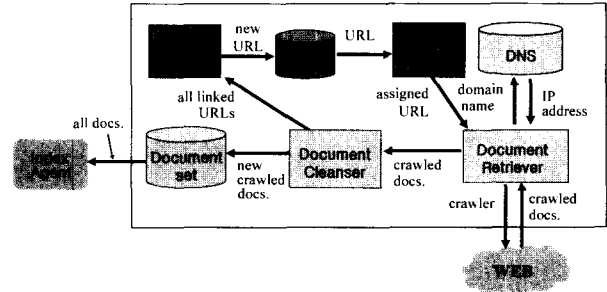


그림 5. RA의 기능 구성도.
Fig. 5. Functional Configuration of RA.

일반적으로 검색로봇의 탐색방법은 깊이우선 탐색, 넓이우선 탐색, 그리고 최적우선 탐색 방법이 있다. 최적우선 탐색은 휴리스틱 방법을 이용하여 다음 URL을 결정한다. 대부분의 검색로봇들은 종자 집합에 있는 페이지들로부터 시작하여 넓이우선 탐색 방법을 수행한다. 그러나 본 논문에서 다루는 전사적 네트워크에서는 계층 3 이하에서만 검색로봇이 작동하고 그 이하 계층에서는 가장 작은 길이의 URL을 가진 링크를 우선으로 검색하는 최적우선 알고리즘의 휴리스틱 기법을 제안한다. 그 이유는 URL의 길이가 작을수록 한 호스트의 최상위 계층의 위치를 나타낼 가능성이 많으므로 제안한 계층적인 구조에 적합하다고 할 수 있다.

5. 구현

5.1 인덱스 DB 설치 및 머지 전략

멀티-에이전트 기반의 웹 마이닝 시스템을 효율적으로 구현하기 위하여 먼저 고려한 요소는 계층적 환경에서의 문서 수집을 위한 효율적인 탐색방법, 각 웹 사이트의 문서를 사용자에게 제공하기 위한 경제적인 인덱싱 방법, 그리고 인덱스 DB의 효과적인 배치 및 머지 방법이다. 이러한 고려요소

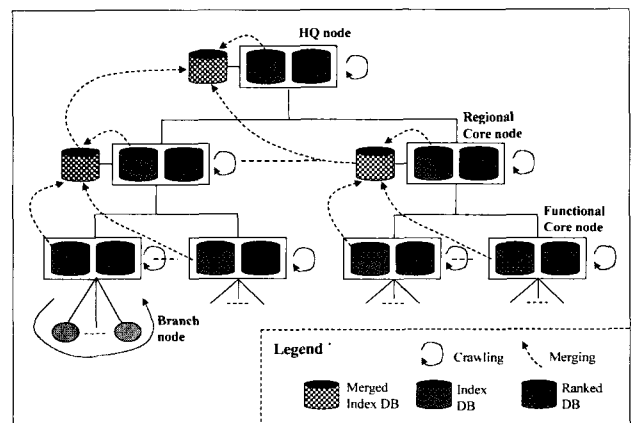


그림 6. 웹 마이닝 시스템의 구현.
Fig. 6. Implementation of Web Mining System.

를 반영하여 구현한 전체적인 구성도가 그림 6에 도시되어 있다.

먼저 검색로봇의 탐색은 중앙노드부터 가지노드까지 모든 계층의 노드에서 수행되나 탐색대상은 계층마다 달라진다. 즉, 기능핵심노드는 자신의 문서뿐만 아니라 자기의 가지노드 문서까지 검색로봇을 이용하여 수집해야 한다. 그러나 중앙노드와 지역핵심노드는 자신의 문서만 탐색하여 수집하고 하위계층 노드의 문서는 탐색하지 않고 버지만 한다. 그리고 지역핵심노드 간에는 관련된 하위 가지노드만 탐색을 통해 문서를 수집해야 되며 서로 간에 로봇을 중복하여 운용하지 않도록 한다.

다음은 검색로봇이 수집한 문서를 사용자에게 제공하기 위하여 DB에 저장하는 인덱싱에 대하여 살펴본다. 인덱싱은 가지노드를 제외한 모든 노드에서 이루어진다. 그러나 위에서 설명한 탐색의 대상이 계층마다 다르므로 인덱싱의 대상 또한 계층에 따라 달라진다. 이러한 방안은 문서의 우선순위 부여시에도 적용되어야 한다.

인덱스 DB의 효과적인 배치 및 버징을 위해 인덱스 DB는 가지노드를 제외한 모든 노드에서 관리되나 버지된 인덱스 DB는 계층에 따라 다르게 관리한다. 즉, 중앙노드와 지역핵심노드에만 두고 기능핵심노드에는 두지 않는다. 다음은 인덱스 DB를 위한 버징 프로시저를 보여준다.

procedure MergeIndex

```

Let I be a merged index DB of current node;
I0 be an index DB of current node;
Ii be an index DB of ith subnode of current node;
where 1 ≤ i ≤ n;
Let U be an union of index DBs;
begin
  for i from 0 to n
    I ← I ∪ Ii;
  end
end
    
```

5.2 랭킹 알고리즘 개선 전략

문서에 점수를 부여하는 랭킹 알고리즘은 기존 PageRank 알고리즘[9]에서 계산된 가중치에 문서위치를 고려하여 계산된 별도의 환경 가중치를 결합함으로써 사용자의 위치까지 고려된 알고리즘을 말한다. 즉, 이 알고리즘은 문서의 하이퍼링크

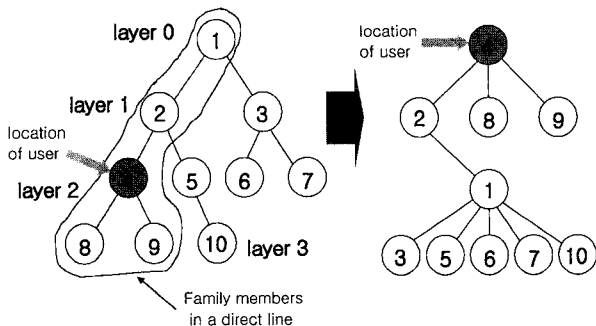


그림 7. 계층적 환경요소를 고려한 노드 4 중심의 가중치 계산을 위한 재배열.

Figure 7. Rearrangement of the whole graph centered on node 4 for computing weighting values considered the hierarchical environment factor.

크 구조와 문서의 실제 위치까지 고려한 알고리즘이다. 이러한 알고리즘이 어떻게 유도되었는지 그림 7에 도시되어 있다.

그림 7의 왼쪽 그래프에서 노드 4에 있는 사용자가 검색을 원한다고 가정한다. 그러면 이 사용자는 자기의 노드에 있는 문서가 제일 앞에 나타나기를 기대할 것이다. 따라서 우리가 제안하는 알고리즘은 기존의 PageRank 알고리즘으로 문서의 콘텐츠를 일차적으로 고려하되 문서의 실제 위치 정보를 같이 고려하여 결합시키자는 것이다. 이러한 경우에 노드 4를 중심으로 부모와 자식과 같은 직계 노드만 허용하여 재정렬하면 그림 7의 우측과 같이 전혀 다른 그래프가 된다. 따라서 계층적 웹 환경에서의 문서의 위치를 고려한 계산된 별도의 환경 가중치인 W_i 는 다음과 같다.

$$W_i = 1 - \left| \frac{l_i - L_p}{L + 1} \right| \quad (2)$$

여기서 l_i 는 노드가 존재하는 계층위치이며, L_p 는 검색을 원하는 사용자의 위치이다. 그리고 L 은 주어진 네트워크의 전체 계층수로서 W_i 의 최대값은 1이 된다.

그리고 그림 7의 우측 그래프를 보면 특수한 경우가 발생하는데 노드 4의 직계를 제외한 노드 3, 5, 6, 7, 10이 노드 1 밑으로 재정렬이 되는데 이들 노드에 대한 환경 가중치를 구한 식은 다음과 같다.

$$W_o = \min(W_i) - \frac{1}{L + 1} \quad (3)$$

여기서 W_o 는 직계가족이 아닌 노드들에 대한 환경 가중치를 말하며 여기서 o 의 의미는 others이다.

최종적으로 계층적인 웹 환경에서 사용자를 고려한 가중치를 구하는 식은 다음과 같다.

$$PR(A)_{improved} = PR(A) \times W_i \quad (4)$$

그림 7에서 나온 각 노드의 가중치를 계층적(노드 1과 4, 노드 1과 9)인 경우와 기능적(노드 2와 3, 노드 4와 5)인 경우에 대하여 가중치를 비교한 그림이 그림 8에 도시되어 있다. 계층적인 경우인 그림 8의 (a)와 (b)의 경우를 보면 계층간의 깊이가 깊어질수록 가중치가 점점 더 많이 차이가 난다는 것을 알 수 있다. 그리고 기능적인 경우인 (c)와 (d)의 경우를 보면 깊이에 많은 차이가 없으며 오히려 계층이 밀어서 내려 갈수록 차이가 더 적어진다는 것을 알 수 있다.

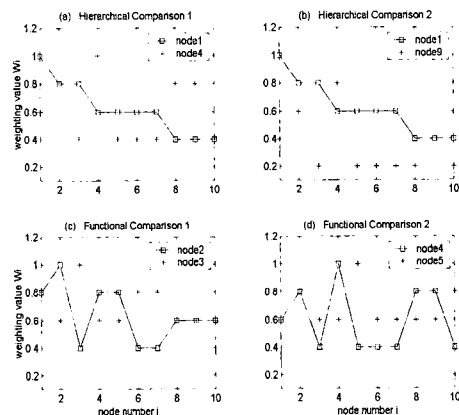


그림 8. 계층적과 기능적인 경우의 가중치 비교.
Figure 8. The comparison of weighting values for hierarchical and functional cases.

6. 결 론

본 논문에서는 전사적 네트워크와 같은 계층적 웹 환경에서 사용자에게 효율적인 정보검색을 지원하는 멀티 에이전트 기반의 정보검색 시스템에 대한 모델링을 제안하였다. 이를 위해 우리는 서로 협력하는 4개의 에이전트와 14개의 에이전트별 프로세스 모듈을 식별하여 그 기능과 구성도를 각각 제시하였다. 제안된 에이전트들의 기능을 주어진 웹 환경에서 구현하기 위하여 경제적인 인덱싱 방법, 인덱스 DB의 효과적인 배치 및 버징 방법을 제안하였다.

또한 웹 기반 환경하에서 사용자의 편의를 고려한 개선된 랭킹 알고리즘을 그래프이론을 이용하여 제안하였다. 마지막으로 우리는 각 계층의 노드에 대한 가중치 변화를 계층적인 경우와 기능적인 경우를 각각 제시하였다. 기능적인 경우에는 계층간의 큰 차이가 발생하지 않았으나 계층적인 경우에는 계층간의 깊이가 깊어질수록 가중치가 점점 더 많이 차이가 난다는 것을 알 수 있었다. 이것은 전사적 네트워크와 같이 논리적으로 계층적인 네트워크에서는 기존의 상용 검색엔진에서 사용하는 랭킹 알고리즘에다 별도의 환경 가중치를 고려하여야 더 사용자에게 유용한 검색결과를 제공해 준다는 것을 의미한다.

참 고 문 헌

- [1] T. Witting, et. al., "ARCHON-A Framework for Intelligent Co-operation," Journal of Intelligent Systems Engineering-Special Issue on Real-time Intelligent Systems in ESPRIT, Vol. 3. No. 3, pp.168-179, 1994.
- [2] Soon-Chul Baek, et. al., "A Framework for Multi-agent Systems in Heterogeneous and Distributed Environment," Journal of KISS(C), Vol. 2. No. 1, pp.24-37, 1996.
- [3] FIPA: FIPA 2000 Specification, 2000.
- [4] Steve Lawrence, C. Lee Giles, "Searching the Web: General and Scientific Information Access," IEEE Communications Magazine, pp.116-122, 1999.
- [5] Feng Guozhen, et. al., "SAInSE: An Intelligent Search Engine Based on WWW Structure Analysis," International Symposium on Parallel and Distributed Processing, pp.1734-1740, 2001.
- [6] Henry Tirri, "Search in Vain: Challenges for Internet Search," IEEE Computer, pp.115-116, 2003.
- [7] Dik L. Lee, Huei Chuang, and Kent Seamons, "Document Ranking and the Vector-Space Model," IEEE Software, pp.67-75, 1997.
- [8] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," Report RJ 10076, 1997.
- [9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Rankings: Bringing Order to the Web," Tech. Report, Computer Science Department, Stanford University, 1998.
- [10] Monika R. Henzinger, "Hyperlink Analysis for the Web," IEEE Internet Computing, pp.5-50, 2001.
- [11] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins, "The Web

and Social Networks," IEEE Computer, pp.32-36, 2002.

- [12] S.Chakrabarti et al., "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text," 7th International World Wide Web Conference, Computer Networks and ISDN Systems, vol. 30, No.1-7, pp.65-74, 1998.
- [13] R. Lempel and S. Moran, "SALSA: The Stochastic Approach for Link-Structure Analysis," ACM Transactions on Information Systems, Vol. 19, No. 2, pp.131-160, 2001.
- [14] Soumen Chakrabarti, Martin van den Berg, and Byron Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," 8th World Wide Web Conference, pp.545-562, 1999.

저 자 소 개



윤희병(Heebyung Yoon)

1983년 : 해군사관학교(이학사)
 1986년 : 연세대학교(공학사)
 1991년 : 미국 해군대학원 전산공학(석사)
 1998년 : 미국 Georgia Institute of Technology 전산공학(박사)
 2002년~현재 : 국방대학교 전산정보학과 조교수

관심분야 : 인공지능, 웹 마이닝, 지능정보시스템, 에이전트시스템, 모바일 웹 서비스, 전술데이터링크

Phone : 02-300-2138

Fax : 02-309-6233

E-mail : hbyoon@kndu.ac.kr



김화수(Hwa-Soo Kim)

1976년 : 해군사관학교(이학사)
 1984년 : 미국 해군대학원 전산공학(석사)
 1990년 : 미국 Case Western Reserve University 전산공학(박사)
 1991년~2002년 : 국방대학교 전산정보학과 교수

2003년~현재 : 아주대학교 정보통신대학원 교수/ 국방부장관 정책자문위원(정보화분야)

관심분야 : S/W 개발비용 산정/분석, S/W유지보수비용 산정/분석, 국방가상현실/시뮬레이션 및 모델링, 최신 정보기술 관리, 인공지능 및 전문가시스템 개발

Phone : 02-319-4080(내선 714)

Fax : 02-754-1182

E-mail : ajhskim@ajou.ac.kr