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[1]
국방대학교, [1]아주대학교 정보통신대학원

Heebyung Yoon, Hwa-Soo Kim[1]
Korea National Defense University,
[1]Ajou University, Graduate School of Information & Communication Technology
E-mail : hbyoon@kndu.ac.kr, [1]aihskim@ajou.ac.kr

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.

1. 서론

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 최신의 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 필요하게 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 등이다.

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 힘들다.

계층적 웹 환경에서의 효율적인 웹 마이닝 시스템을 설계하기 위해서는 웹 사이트마다 시스템을 각각 운용하는 것보다 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.

2. 관련 연구

2.1 멀티에이전트 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. 1996년에는 에이전트 관련 최초의 세계적인 표준화 기구인 FIPA[3]에서 제안한 AP가 있다.

2.2 검색엔진

검색엔진은 크게 사용자 입장에서의 분류와 세대별 분류로 구분할 수 있다. 사용자 입장에서는 주제별 검색엔진, 단어별 검색엔진, 메타 검색엔진으로 분류된다. 세대별은 세대 전 검색엔진, 1세대 검색엔진, 2세대 검색엔진, 3세대 검색엔진으로 분류된다[4].

1세대 검색엔진은 웹 로봇이 인터넷을 돌아다니며 자동으로 웹 문서를 수집해 오는 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. 3 세대 검색엔진은 현재 개발 중인 엔진으로 구문 분석을 할 수 있다는 것이 2세대와 큰 차이점이다[5].

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 지역핵심노드, 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.

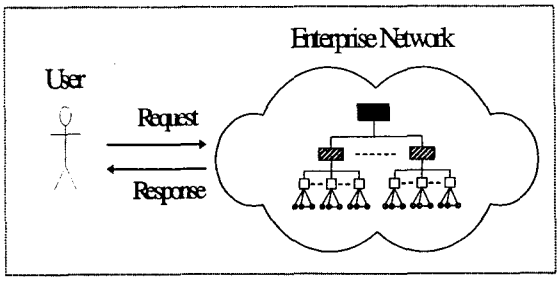

그림 1 계층적 웹 환경

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 고려요소를 계층적인 웹 환경 측면과 정보검색 시스템의 기능 측면 등 두 가지 측면에 살펴볼 수 있다.

먼저 계층적인 웹 환경 측면에서는 각각의 독립적인 기능들을 수행하는 에이전트들을 어디에 배치할 것인가와 인덱스 DB를 어느 노드에 어느 수준까지로 배치할 것인가가 주요한 고려요소가 된다. 그리고 정보검색 시스템의 기능 측면에서는 에이전트의 분류, 에이전트의 기능 식별, 그리고 검색된 문서에 우선순위를 부여하는 랭킹알고리즘 등이 주요한 고려요소가 된다.

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 분류하였다. User Interface Agent(UIA), Merge Agent(MA), Index Agent (IA), 그리고 Robot Agent(RA)이다.

UIA는 사용자의 요청 정보를 분석하고 검색결과를 사용자에게 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-에

이전트 기반의 웹 마이닝 시스템 구성도가 그림 2 에 도시되어 있다. 참고로 이 그림에서는 웹 마이닝 시스템에 대한 에이전트 기능을 중심으로 하여 분류하였으며 에이전트간 정보교환을 위한 통신프로토콜 및 구조는 생략하였다.

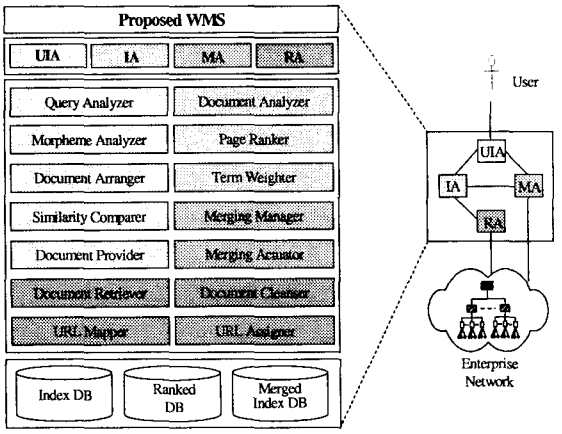

그림 2 제안한 멀티-에이전트 기반의 웹 마이닝 시스템 구성도

그림 2에서 제안한 웹 마이닝 시스템은 총 세 개의 계층으로 구성되어 있다. 제일 상위 계층은 웹 마이닝 시스템 계층이며 그 아래가 에이전트 계층이고 다음이 프로세스 모듈계층이다. 에이전트 계층에는 네 개의 에이전트가, 프로세스 모듈 계층에는 14개의 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 다음과 같은 다섯 개의 모듈로 구분할 수 있다. Query Analyzer는 사용자의 질의를 받아 분석하며 형태소 분석을 위하여 질의 단어를 Morpheme Analyzer로 제공한다. Morpheme Analyzer는 사용자의 질의 단어를 분석하여 불필요한 단어를 제거하며 어간 정제과정을 거쳐 합성어를 제공한다. Similarity Comparer는 사용자의 질의어가 합성어로 입력되면 Index DB의 문서와 유사도를 비교하여 검색된 문서의 식별번호를 제공한다. Document Arranger는 문서의 ID로 우선순위가 부여된 문서를 찾아내며 사용자의 IP 주소를 고려하여 최종적인 문서를 우선순위별로 제공한다. Document Provider는 사용자에게 최종적으로 순위화된 문서를 URL, 문서내용, 유사도 등과 함께 제공한다.

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 보낸다. Term Weighter는 Vector Space Model의 tf x idf를 이용하여 색인어의 가중치를 계산하며 문서의 ID 순서로 가중치와 인덱스 단어를 포함하여 인덱스 DB에 저장한다. Page Ranker는 문서의 하이퍼링크 구조와 위치정보를 이용하여 각 문서에 점수를 부여한다.

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

일반적으로 웹 로봇의 탐색방법은 깊이우선 탐색, 넓이우선 탐색, 그리고 최적우선 탐색 방법이 있다. 그러나 본 논문에서 다루는 전사적 네트워크에서는 계층 3 이하에서만 웹 로봇이 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

우리는 위에서 제안한 멀티 에이전트 기반의 웹 마이닝 시스템을 구현하기 위하여 다음의 세 가지 사항을 고려하였다. 1) 계층적 환경에서의 문서 수집을 위한 효율적인 탐색방법, 2) 각 웹 사이트의 문서를 사용자에게 제공하기 위한 경제적인 인덱싱 방법 및 인덱스 DB의 효과적인 배치와 머징 방법 그리고 3) 웹 환경 요소를 고려한 개선된 랭킹 알고리즘이다.

먼저 웹 로봇의 탐색은 중앙노드부터 가지노드까지 모든 계층의 노드에서 수행되나 탐색대상은 계층마다 달라진다. 즉, 기능핵심노드는 자신의 문서뿐만 아니라 자기의 가지노드 문서까

지 웹 로봇을 이용하여 수집해야 한다. 그러나 중앙노드와 지역핵심노드는 자신의 문서만 탐색하여 수집하고 하위 계층노드의 문서는 탐색하지 않고 머지만 한다. 그리고 지역핵심노드 간에는 관련된 하위 가지노드만 탐색을 통해 문서를 수집해야 되며 서로 간에 로봇을 중복하여 운용하지 않도록 한다.

다음은 웹 로봇이 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. 인덱스 DB의 효과적인 배치와 머징을 위해 인덱스 DB는 가지노드를 제외한 모든 노드에서 관리되나 Merged Index DB는 계층에 따라 다르게 관리한다. 즉, 중앙노드와 지역핵심노드에만 두고 기능핵심노드에는 두지 않는다. 다음은 Merged Index DB를 수행하는 절차를 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 ∪ be an union of index DBs;
    begin
        for i from 0 to n
        I ← I ∪ Ii;
    end
  end
```

마지막으로 문서에 점수를 부여하는 랭킹 알고리즘은 기존 PageRank 알고리즘[7]에서 계산된 가중치에 문서위치를 고려되어 계산된 별도의 환경 가중치를 결합함으로써 이루어진다. 계층적 웹 환경에서의 문서의 위치를 고려한 계산된 별도의 환경 가중치인 W_i는 다음과 같다.

$$W_i = 1 - \frac{l_i - L_p}{L + 1} \qquad (1)$$

여기서 ℓ_i 는 노드가 존재하는 계층위치이며, L_p는 검색을 원하는 사용자의 위치이다. 그리고 L은 주어진 네트워크의 전체 계층수로서 W_i의 최대값은 1이 된다.

수식 1을 이용한 사용자의 검색위치에 따른 가중치 변화가 그림 3에 도시되어 있다. 가중치 계산에 사용한 예제 그래프는 전체 4개의 계층에 10개의 노드가 있으며 노드 1이 최상위 노드이다(그림 3 좌측). 이 그림에서 보면 계층적인 경우와 기능적인 경우 모두 검색 위치(노드의 위치)에 따라 가중치가 크게 차이가 난다는 것을 알 수 있다.

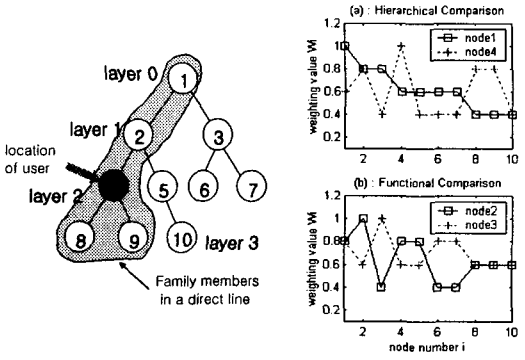

그림 3 검색위치에 따른 가중치 변화

5. 결론

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 4 개의 에이전트와 14개의 에이전트별 모듈을 식별하여 그 기능을 각각 제시하였다. 제안된 에이전트들의 기능을 주어진 웹 환경에서 구현하기 위하여 웹 로봇의 탐색방법, 경제적인 인덱싱 방법, index DB의 효과적인 배치 및 머징방법, 그리고 웹 환경 요소를 고려한 개선된 랭킹 알고리즘을 제안하였다.

6. 참고문헌

[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] 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.

[5] Henry Tirri, "Search in Vain: Challenges for Internet Search," IEEE Computer, pp.115-116, 2003.