1999 年度 大韓電子工學會 夏季綜合學術大會 論文集 第 22 卷 第 1 號 99/6

사용자 정보와 자동 문서 분류를 이용한 웹 에이전트의 설계

°이 승 원, 권 영 훈, 류 제, 한 광 록
호서대학교 컴퓨터공학과
lhonesty@mail.hoseo.ac.kr
krhan@office.hoseo.ac.kr

Design of Web Agent Using User Profile and Automatic Document Categorization

Lee Seung Won, Kwon Young Hoon, Ryu Je, Han Kwang Rok
Dept of Computer Engineering, Hoseo University

Abstract

WWW is an important method for retrieving or providing informations. Not only the amount of information but also it is widely located on the web, it is difficult for users to get or search information. Furthermore, to use search engine is also inconvenient, because it just uses a keyword without concerning a user's interest. At this point, we propose a design of web agent that uses the automatic document categorization system and user's profile concerning with a user's interest, so the agent can actively provide a information.

1. 서론

웹 상에서 정보를 찾고자 하는 대다수의 사용자들이 선택하는 검색 방법은 검색엔진을 이용하여 찾고자 하는 키워드를 입력하여 제시된 결과 중에서 자신이 원하는 정보를 찾는 것이다. 그러나 이러한 방법은 사용자의 관심도는 전혀 반영될 수 없으며 보편적인 의미의 키워드일 경우, 검색결과 중에서 원하는 정보를 찾는 것 또한 쉬운 일이 아니다. 이에 웹 상에서 존재하는 문서들을 대상으로 자동 분서분류를 적용하며 정의된 category 에 따라 분류를 시도하고 사용자의 관심도를 반영한 Profile을 구축하여 기존 또는 새롭게 획득한 문서들 중에서 사용자가 관심을 표명한 category 내의 문서를 능동적으로 제시하는 웹 에이전트를 제시하고자 한다[8]. 이러한 웹 에이전트는 사용자가 입력한 키워드들에 대한 수동적인 정보검색이 아니라 문서분류기술을 이용하여 보다 능동적으로 웹 상의 정보를 사용자에게 제시함을 그 목적으로 한다. 본 논문에서는 웹 상에 존재하는 문서들을 획득하는 Web Robot Agent, Web Robot Agent에 의하여 얻어진 문서들을 정의된 Category 에 따라서 자동으로 분류하는 Document Classify Agent, 그리고 각 Category 의 정보들을 사용자에게 제공하며 사용자의 Feed Back을 처리하는 User Interface Agent 로 구성되어 있다.

2장에서는 기존의 검색방식을 개선하고자 연구되고 있는 관련연구에 대하여 기술하고 3장에서는 본 논문에서 제안하고 구현한 각 Agent 들에 대하여 상술하며 4장에서 실험방법을 기술하고 5장에서 결론을 내렸다

2. 관련연구

사용자 정보를 이용하고자 하는 것은 기존의 검색방식이 획일적으로 사용자가 입력한 키워드들과 문서 DB와의 검색결과만을 제시하는 단점을 해결하고자 하는 것이다.

2.1 인터넷 정보 여과 에이전트

정보 여과(information filtering)는 기본적으로 끊임없이 유입되는 정보 중에서 필요한 것이 무엇이고 필요없는 것이 무엇인지를 판단하여 필요하지 않은 것은 무시하는 개념이다[11]. 현재 진행되고 있는 많은 수의 연구에서는 정보를 제공받을 사용자의 Profile을 이용하며, 사용자의 Profile 에는 사용자가 관심을 가지는 사항에 대한 정보가 포함되고 정보 여과의 과정은 email 등의 정보와 같은 정보 스트림을 사용자의 profile과 비교하여 관심이 있는 정보만을 filtering 하여 사용자는 여과된 정보만 볼 수 있게 한다. 사용자는 이러한 정보에 대하여 그것이 실제 자신이 원하는 정보였는가에 대한 가부 여부를 알려주게 되는데 이를 relavance feedback[11]이라 하며 이러한 과정을 거쳐 사용자 Profile을 재구성할 수 있다.

현재 웹 상에서 다수의 정보 여과 에이전트들이 연구용 또는 상용으로 제시되었다. WebFileter[5], Webcatcher[6]등은 웹 상의 문서 여과 에이전트들이며 NewsHound[7], PointCast 등은 상용 뉴스 여과 에이전트들이다.

2.2 정보 통합 에이전트

정보 통합 에이전트(information integration agent)는 인터넷에서 제공되는 다수의 정보 사이트에서 사용자가 원하는 정보를 추출하여 하나의 형태로 제공하는 기능을 수행한다[11]. 이러한 에이전트의 필요성은 정보 분포상의 광대함에 대한 어려움을 줄여 줄 수 있다는 데 있다. 정보 통합 에이전트의 예로는 BarginFinder를 들 수 있다.

3. 에이전트의 구조

그림[1]은 본 논문에서 제안하는 시스템의 전체 흐름도이다[2]. 제 2장에서 서술했던 기존의 관련연구와의 공통점은 웹 상에서 획득한 문서에 대하여 사용자의 관심도를 이용한다는 점이며 차이점은 사용자가 검색하고자 입력한 키워드들에 대하여 단순히 Profile을 이용한 문서의 Filtering을 수행하는 것이 아니라 자동 문서 분류 과정을 통하여 구축된 데이터를 사용자의 관심도에 따라 능동적으로 정보를 제공한다는 점이다.

3.1 로봇 에이전트

로봇 에이전트란 웹서버를 순회하며 각 홈페이지에 있는 수많은 정보를 수집하는 프로그램이다. 본 논문에서는 로봇에이전트를 이용하여 문서를 수집하고 분류하는 방식을 택하고 있다[13]. 로봇 에이전트를 사용할 경우

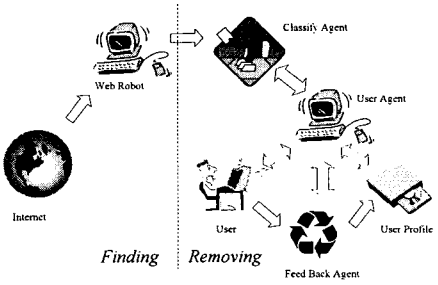

그림 1 시스템 흐름도

문서획득에 소요되는 시간을 절약할 수 있으며, 또한 새로운 URL의 검색 및 새로운 문서획득이 용이하다. 그러나 로봇 에이전트는 네트워크에 과도한 부하를 줄 수 있기 때문에, 본 논문에서는 로봇 에이전트가 수행되는 시간을 최소한으로 하면서 많은 양의 문서를 수집하도록 하는데 주안점을 두었다.

본 논문에서는 로봇 에이전트의 URL 데이터 베이스로서 Hash Table을 사용하여, URL검색 및 새로운 URL의 추가, 삭제 등에 소요되는 시간을 최소화 함으로써, 로봇 에이전트의 효율을 높이는 방식을 채택하였다[12]. 그림[2]는 로봇 에이전트의 전체 구조를 나타낸다.

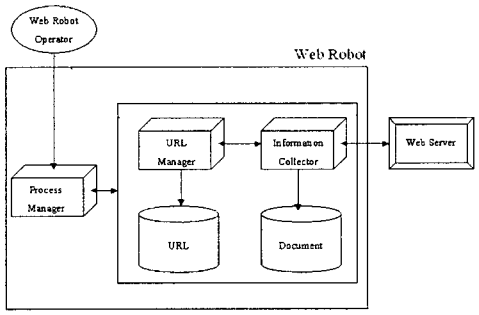

그림 2 Web Robot 구성도

우선 Process Manager는 로봇 에이전트의 전체 동작을 조절한다. 다음으로 URL Manager는 주어진 URL을 시작으로 하여 Information Collector를 통하여 얻어진 URL들을 Hash Table과 비교하고, 새로운 URL일 경우 Table에 추가하며, 검색할 사이트의 URL은 Queue에 넣어둔다. URL의 검색 순서는 URL의 Hyper Link 의 구조가 Tree의 형태인 점을 감안하여 넓이 우선 탐색(Breadth-First Search)방식을 사용한다. Information Collector는 검색한 URL에서 얻어진 문서는 Document DB에 저장하고, 문서에 링크된 URL들은 URL Manager에 전달하는 역할을 수행한다. 로봇 에이전트는 Hash Table을 사용함으로써 URL 관리를 효율적으로

함으로써 로봇 에이전트의 수행 시간을 줄이는 특징을 가지고 있다.

3.2 자동 문서 분류(Classify Agent)

문서의 분류란 정해진 분류체계 하에서 분류하고자 하는 각 문헌들을 가장 적합한 Category에 배정함으로써 문헌을 집단화하는 작업이며[3], 문서의 수동 분류에 대한 한계로 인하여 현재 활발한 연구가 진행중이다[1][4]. 본 논문에서는 웹 상의 문서들을 8개의 최상위 범주와 각 하위범주로 문서의 분류를 시도한다[표1].

표 1 최상위 분류체계

1	컴퓨터와 인터넷	5	건강과 의학
2	사회생활	6	레크리에이션과 스포츠
3	경제생활	7	과학과 공학
4	문화생활	8	교육

문서의 자동 분류에 사용되는 방법에는 단순한 단어의 매칭을 이용하는 방법과 확률을 이용하는 방법, 통계적인 기법에 인공지능 기법을 이용하는 방법, Neural Network를 이용하는 방법 등이 있으며 본 논문에서는 일반적인 정보검색에서 사용되는 Vector Space Model의 IDF(Inverse Document Frequency)를 변형한 ICF(Inverted Category Frequency)를 이용하여 문서의 분류 정확성 향상을 시도한다[9]. 문서 분류에 있어 ICF의 장점은 다음과 같다. D1,D2,D3,D4 네 개의 문서가 있고 (D1,D3),(D2,D4) 가 각각 category를 이루고 있다고 가정할 때 키워드 W1은 D1,D2에 나타나고 W2는 D1,D3에 나타난다면 W2의 가중치를 더 크게 주어야만 문서의 분류에 도움을 줄 수 있다. 그러나 일반적인 IDF상에서는 두 개의 키워드 W1,W2의 가중치가 같은 문제점을 가지고 있다.

본 논문에서는 문서 분류의 효율성을 위하여 두 개의 백 데이터를 이용한다[10]. 문서 분류의 기초 자료를 위하여 분류체계에 따른 범주별 백 데이터와 이러한 범주별 백 데이터를 통합한 통합 백 데이터가 그것이다. 통합 백 데이터는 그 정확성을 위하여 대규모의 문헌집합에서 구축되어야 하며 색인 용어의 선택의 문제점을 극복하기 위하여 기존에 구축된 시소러스를 이용한다. 하나의 문서에서 추출된 키워드들과 이렇게 구축된 통합 백 데이터와의 비교를 통하여 분류여부를 결정하게 된다. 즉, 문서에서 추출된 키워드가 통합 백 데이터에 존재할 경우 그 키워드의 가중치를 증가시키게 된다. 즉 가중치는 키워드의 문서 내와 통합 백 데이터의 존재여부에 비례하며 각 키워드의 가중치 총합이 어느 임계치 이하일 경우는 그 문서에 대한 분류를 중지하게 된다. 이렇게 통합 백 데이터와의 1차 비교 이후에 각 범주별 가중치를 구하게 되는 데 각 범주별 백 데이터와의 ICF 값을 구하여 가장 큰 유사도를 보인 범주로서 문서가

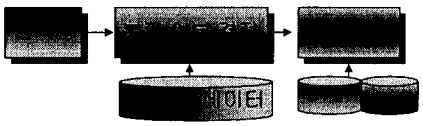

그림 3 문서의 분류 과정

분류가 된다. Vector Space Model 에서의 유사도 계산은 분류하려는 문서와 구축된 범주별 기초 데이터들을 색인어들의 벡터로 표현하고, 두 벡터 사이의 유사도를 계산하여 가장 큰 값을 가지는 범주로 문서를 분류하는 방법이다. 분류하려는 문서에 대한 벡터를 D, 어느 범주의 벡터를 Ci 라고 한다면 두 벡터 D와 Ci 간의 유사도는 다음과 같이 계산된다.

$$Sim(D,Ci) = \frac{D \bullet Ci}{|D| \ |Ci|}$$

또한 벡터 표현에서 정확도를 높이기 위하여 각 색인어에 가중치를 부여하는데 정보검색에서 이용하는 IDF를 변형한 ICF에서는 다음과 같이 계산된다[3].

$$ICF_i = \log_2 \frac{N}{n_i} + 1$$

n_i = the total number of category include term i
N = the number of Category

ICF를 이용한 각 색인어의 가중치는 다음과 같다.

$$W_{ij} = freq_{ij} * ICF_i$$

W_{ij} : j 번째 문서에서 i 색인어의 가중치
$freq_{ij}$: j 번째 문서에서 i 색인어의 빈도수

이것은 하나의 범주내의 색인어가 다른 범주에서의 빈도수가 적고 그러한 색인어를 많이 포함하는 문서에 대하여 그 범주에 포함시키는 것이다.

3.2.1 하위 범주에 대한 문서 분류

일반적인 경우 이러한 범주는 분류체계가 수평적이지 않다는 문제가 있다. 즉 하나의 상위 범주는 또 다른 하위 범주를 포함하고 있다. 이러한 계층적 분류에서는 수평적인 분류체계에서와는 달리, 범주의 깊이를 정의하는 문제와 하나의 상위 범주로 분류된 문서를 현재보다 하위의 범주에 어떻게 분류하는 가에 대한 문제가 발생한다. 본 논문에서는 그 목적이 문서의 분류에 있지 않고 분류된 문서의 정보를 사용자의 관심도에 따라 제공하는 것에 있으므로 범주의 넓이와 깊이를 모두 고정적으로 가정한다. 일반적인 트리 구조를 가지는 계층적 분류체계에서는 한 문서에 대하여 분류를 시도하여 만약 아

무런 범주에도 문서가 포함되지 않았을 경우는 현재의 범주로 분류 실패를 의미하며 분류가 되었을 때는 그 범주의 모든 하위 범주에 대하여 유사도를 계산하여 가장 높은 유사도를 갖는 범주를 선택하도록 한다. 여기에서 특정 임계치를 정하여 분류의 오류를 줄이는 방법을 쓸 수도 있으며 만약 하위 범주에 대한 유사도 계산결과 임의의 임계치를 넘는 결과가 없을 경우는 현재의 범주를 최하위 범주로 선택한다. 또한 한 문서에 대하여 범주가 선택되었을 때 그 상위 범주에 대하여서는 하위의 개념은 상위의 개념을 포함함으로 선택된 범주의 상위 범주 모두에 포함시키도록 한다.

3.3 사용자 정보 관리(User Interface Agent)

사용자의 관심도를 반영하고 있는 Profile의 갱신 방법은 일반적으로 사용자가 자신의 관심도를 표현하는 Supervised Learning 방법과 사용자의 행동을 모니터링 함으로써 동적으로 Profile을 갱신하는 Unsupervised Learning 방법이 있으며 본 논문에서는 전자의 방법을 이용하도록 한다. 이렇게 구축/갱신된 Profile을 참조하여 위의 문서 분류를 통하여 얻어진 문서 정보들을 사용자에게 제공하게 된다.

4. 실험

본 논문에서는 3장에서 기술한 분류체계를 이용하고, 문서에서의 색인어 추출과 분류체계에 따른 통합 및 범주별 백 데이터 구축과 ICF를 이용한 문서의 분류에 선행 연구로 구축된 50,000 명사 표제어의 상위-하위 개념 분류와 7,000 개의 동의어에 대한 정보가 수록된 시소러스와 ETRI의 형태소분석기와 자동태거를 사용하였다. 색인 및 중복제거와 카테고리화를 거친 데이터의 구조는 그림[4]와 같으며 중앙일보 경제란 기사를 이용하여 경제와 스포츠 카테고리와의 유사도 결과는 표[2]와 같다.

	ID	word	location	fre	weight
	390	조건인수	경제9.txt	1	.3010300099B5
	391	삼성	경제10.txt	7	2.107209920883
	392	대우	경제10.txt	5	1.505149960518
	393	빅딜	경제10.txt	11	3.311329841614
	394	자동차	경제10.txt	5	1.505149960518

그림4 기초 데이터의 예

표 2 경제 기사와의 유사도

경제 카테고리	20.47004
스포츠 카테고리	0.60206

5. 결론 및 향후과제

본 논문에서는 획일적인 현재의 정보검색의 단점을 극복하고 나아가 사용자의 관심도에 따른 문서 정보의 동적 제공을 위하여 ICF를 문서의 자동 분류와 이를 위한 방법들을 제안하였다. 이것은 정보를 검색하기 위한 사용자의 행동 단계를 줄일 수 있으며 기존의 수동적에서 동적으로의 변환에 그 의의가 있다고 하겠다.

앞으로 분류의 성능을 더욱 높이기 위하여 각 단계에서 택한 임계치의 설정문제와 고정된 범주 체계에서 오는 분류실패를 줄이는 문제, 그리고 사용자의 관심도를 반영하는 Profile의 효과적인 구축과 동적 갱신에 관한 다각적인 연구가 필요하며, 현재 Issue 가 되고 있는 최저 가격 정보 제공에 관한 연구도 진행되어야 할 것이다.

참고문헌

[1] Rainer Hoch, "Using IR Techniques for text Classification in Document Analysis, SIGIR, 1994

[2] Takuo Nakashima, "Information Filtering for the Newspaper", IEEE, 1997

[3] William B. Frakes, "Information Retrieval", PrenticeHall, 1992

[4] David D.Lewis, "Evaluating and Optimizing Autonomous Text Classification Systems", SIGIR,1995

[5] WebFilter, http://ils.unc.edu/webfilter.

[6] Webcatcher,http://plum.tuc.noao.edu/webcatcher/webcatcher.html

[7] NewsHound, http://www.sjmercury.com/hound.htm

[8] 백혜정, 박영택, 윤석환, "사용자 관심도를 이용한 웹 에이전트",정보처리학회지, 1997

[9] 조광제, 김준태, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류", 정보과학회 춘계학술발표집, 1997

[10] 조태호,"동적 정보 키워드 선택에 의한 텍스트 범주화", 정보처리학회 춘계학술발표집, 1999

[11] 최중민, "인터넷 정보 가공을 위한 에이전트" 정보처리학회지, 1997

[12] Gun-Woo Nam, "Dynamic Management of URL Based on Object-Oriented Paradigm",IEEE,1998

[13] Lee Sung-Min, "A New on Demand Service System based on Robot Agent",IEEE,1998