웹 도큐먼트 기반 연관 지식 추출 기법 : 생명정보분야에의 적용

Web Document-based Associate Knowledge Extraction Method : Applying to Bioinformatics

문 현 정* 김 교 정**
Hyun-Jeong Moon Kio-Chung Kim

요 약

본 논문에서는 웹 도큐먼트로부터 사용자의 관심과 선호도를 반영하는 지식을 자동으로 확장 탐색하고 추출하기 위한 연관지식 추출 기법을 제시한다. 사용자의 학습의도를 내포한 중심어와 연관된 정보를 예제 도큐먼트로부터 탐색 추출하기 위하여 연관 규칙 탐색 데이터마이닝 기법을 웹 도큐먼트상의 연관 객체 추출에 적용한다. 또한 추출된 연관 정보들의 가중치 부여를 위하여 연관 태그 블록 기반 가중치 기법을 제시한다. 본 논문에서 제시된 연관 지식 추출 기법을 생명정보학 분야에 적용하여 의미적으로 연관성 있는 지식 추출 실험을 수행한 결과 매우 높은 정확성을 보이는 것으로 나타났다.

Abstract

In this paper, we develop associate knowledge extraction method for finding and expanding user preference knowledge automatically from web document database. To reflect user interest or preferences, agent explores and extracts relevant information to central term involving the intent of users from the example documents. To do so, we apply association rule exploration data-mining method to the extraction of the relevant objects in the web documents. Also, to give the weighted-value to the extracted and relevant information, we present associate tag block-based weighting method. We applied to bioinformatics above associate knowledge extraction method to find related keywords.

1. 서 론

최근 인터넷의 발달로 인한 전자 정보량의 폭발적으로 증가로 인해서 사용자가 필요로 하는 정보를 찾기 위한 시간과 노력이 증가하는 정보 과잉(information overload) 상태가 발생하고 이를 해결하기 위하여 지능적 정보 에이전트(intelligent information agent)시스템이 제시되었다. 지능적 정보 에이전트는 사용자의 습관과 요구에 대하여 높은 적응성(adaptability)을 보여야 하며, 이를 위한 일련의 작업들을 자율성(autonomy)을 갖고 수행하고, 새로운 지식(knowledge)을 탐색하며, 경험(experience)을 통하여 자신의 성능을 향상시킬 수 있어야 한다[1].

따라서 지능적 정보 에이전트는 도큐먼트가 사용자의 관심과 선호도에 부합하는 데이터를 포함하고 있는지 아닌지 만을 구분하는 것으로는 충분하지 않다. 따라서 에이전트 자신의 적응성을 바탕으로 새로운 지식(knowledge)을 탐색 혹은 발견하여 제공하는 지능적인 가이드(intelligent guide)로서의 역할이 요구되어 진다[2,3].

질의-기반 정보 검색 시스템(query-based information retrieval system)은 대량의 도큐먼트들로부터 사용자 질의를 포함한 도큐먼트를 검색하여 제공하는 시스템이다[4]. 일반적으로 Yahoo등과 같은 웹 검색 엔진들과 많은 데이터베이스 시스템들이 질의

* 준회원 : 숙명여자대학교 대학원 컴퓨터과학과 박사과정
hjmoon@cs.sookmyung.ac.kr
** 비회원 : 숙명여자대학교 정보과학부 교수(멀티미디어 전공)
kiochkim@sookmyung.ac.kr

기반 검색을 기반으로 하고 있다[3].

질의-기반 혹은 키워드-기반(keyword-based) 정보 검색은 매우 유용하나 도큐먼트내의 유용한 확장 지식을 탐색하는데 제한적이다[5]. 이를 위하여 의미론적 문맥(semantic context)을 고려한 질의 확장(query expansion)기법을 적용하여 사용자에게 연관된 지식(knowledge)을 제공하는 다양한 기법들이 제시되고 있다[6-8].

지능적 정보 에이전트는 사용자의 정보 요구(혹은 질의)와 선호도에 대한 학습의 결과로 사용자 프로파일(user profile)을 자신의 지식 베이스(knowledge base)의 일부로 구축한다. 사용자 프로파일 모델(user profile model)로는 InfoFinder [9]에서와 같은 불 서치 트리(Boolean search tree) 모델과 키워드-벡터(keyword-vector) 모델 그리고 확률 모델등이 있으며 이중 키워드-벡터 모델링 기법이 웹과 같은 방대한 정보 공간에서 사용자가 원하는 정보를 추출하기 위하여 가장 널리 쓰이고 있다[10-14].

본 논문에서는 사용자 중심어와 연관된 지식을 예제 도큐먼트로부터 추출하는 웹 도큐먼트 기반 연관 지식 추출 기법을 제시한다. 이 기법을 이용하여 지능적 정보 에이전트는 도메인에 대한 사전 지식(prior knowledge) 없이 사용자가 제시한 예제 도큐먼트와 학습의도 중심어로부터 연관 지식을 탐색하여 제시한다. 제시된 연관 지식은 사용자 질의확장에 적용되거나 혹은 사용자 프로파일로 저장되어 사용자의 선호도에 대한 지식으로 구축될 수 있다. 실험 결과에 따르면 최소한의 의미 분석(semantic analysis) 사용에도 불구하고 데이터마이닝(data mining) 기법[15]과 로컬 컨택스트(local context)[8] 개념을 적용하여 중심어와 연관된 도큐먼트 객체를 효과적으로 추출함을 알 수 있다.

본 연구에서 제시된 지능적 정보 에이전트는 생명정보학(bioinformatics)분야에 적용하여 개발 및 실험하였다. 최근 생명정보학 관련 정보처리 분야에서는 관계성 탐색과 추출 그리고 키워드 탐색 이 두 분야가 가장 집중적인 관심과 연구가 진행 되고 있다[3]. 생명정보학 분야는 분자생물학(molecular biology)이나 의학 분야와 더불어 매우 빠른 변화가 있는 분야이며, 매우 복잡하고 많은 동의어(synonym)들이 존재하며 또한 새롭게 생겨나고 있다. 이러한 상황에서 질의-기반 정보검색의 연관 정보를 탐색하고 지식을 확장하는데 따른 제한점이 큰 문제로 대두되고 있다. 따라서 생명정보학 분야에 있어서 연관된 지식을 탐색하여 제공함으로써 사용자의 질의 확장(query expansion) 혹은 주제에 대한 이해(understanding)를 돕기 위하여 지능적인 가이드로서의 지능적 정보 에이전트는 매우 유용한 도구이다. 본 논문에서는 최소의 의미 분석과 단순한 통계적 측정 그리고 데이터마이닝 기법을 적용하여 효과적인 연관 정보 추출 결과를 얻을 수 있었다.

2. 관련연구

2.1 개인화된 지능적 정보 에이전트(Personalized Intelligent Information Agent)

정보의 종류가 다양해지고 정보의 양이 증가할수록 사용자가 필요로 하는 정보를 찾기 위한 시간과 노력이 증가하는 정보과잉(information overload) 상태가 발생하고 이를 해결하기 위하여 지능적 정보 에이전트 시스템이 제시되고 있다[16].

개인화된 지능적 정보 에이전트는 월드와이드 웹과 같은 방대한 정보 집합 속에서 사용자의 정보 요구(information needs) 혹은 관심(interest), 선호도(preference)에 대한 관련 정보(relevant information)를 제시함으로써 사용자를 돕도록 의도되어진 지능적인 시스템이다[2].

따라서 개인화된 지능적 정보 에이전트는 도큐먼트가 사용자의 관심과 선호도에 맞는 데이터를 포함하고 있는지 아닌지 만을 구분하는 것으로는 충분하지 않으며, 에이전트 자신의 적응성을 바탕으로 새로운 지식(knowledge)을 탐색 혹은 발견하여 제공하는 지능적인 가이드(intelligent guide)로

서의 역할이 요구되어 진다[2,3].

2.2 에이전트의 사용자 학습(User Preference Agent Learning)

지능적인 정보 에이전트의 사용자 선호도 학습을 위한 방법으로는 명시적 피드백(explicit feed back) 방법, 암시적 피드백(implicit feedback) 방법 그리고 화이트보드(whiteboard)방법과 같이 크게 세가지로 나눌 수 있다[17].

명시적 피드백(explicit feedback)은 사용자가 에이전트에게 자신의 의사를 명확히 전달하는 기법으로 대표적으로 연관성 피드백(relevance feedback) 기법을 들 수 있다. 즉, 사용자는 정보에 대하여 연관 혹은 비연관의 평가를 내리고 그 결과를 에이전트에게 전달하여 자신의 의도를 학습하도록 한다[18].

암시적 피드백(implicit feedback)은 사용자의 행위를 관찰 함으로써 관심도를 찾아내는 방법이다. 대표적으로 Personal Web Watcher와 Letizia의 경우 사용자의 브라우징 패턴을 관찰하여 하이퍼링크를 추천하거나 페이지를 보여주는 예를 들 수 있다[19].

화이트보드(whiteboard)방법은 에이전트 학습 기법이라기 보다 사용자가 직접 자신의 프로파일을 조작하는 방법으로서 사용자 프로파일 구축을 위한 에이전트 학습과정이 필요 없고 명확하게 사용자의 선호도를 반영할 수 있는 방법이다.

사용자 프로파일 구축을 위하여 가장 널리 적용되고 있는 기법은 명시적 피드백 기법인 연관성 피드백(relevance feedback) 이다[14].

다음에 제시된 Rocchio 알고리듬<식1>은 연관성 피드백 기법들중 일반적으로 모든 도메인에서 좋은 결과를 보이는 기법으로 알려져 있다[18].

사용자는 각 도큐먼트가 그들의 정보요구에 관련되는지 아닌지를 판단하여 지능적 정보 에이전트에게 제시하면, 제시된 도큐먼트들을 학습예제로 사용자 선호도 프로파일을 구축하는 기법이다.

$$Q_1 = Q_0 + \beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2} \quad (1)$$

Q_0 : 초기 질의 벡터
Q_1 : 확장 질의 벡터
R_i : 관련문서 i의 벡터
S_i : 비관련 문서 i의 벡터
n_1, n_2 : 선택된 관련/비관련 문서의 수
β, γ : 관련/비관련 용어의 중요도 조정 가중치

즉 양성(positive) 예제의 속성은 사용자 프로파일에 추가되고 음성(negative) 예제의 속성은 사용자 프로파일로부터 제거되는 과정을 거쳐 사용자 학습을 수행한다.

2.3 사용자 프로파일 모델링(User Profile Modeling)

지능적 정보 에이전트의 사용자 프로파일은 그 응용분야와 목적에 따른 사용자 모델링의 결과로 다양한 형식으로 구축된다. 그 공통적인 목적은 사용자의 선호도를 효과적으로 해석하여 저장하고, 개인화 정보 제공 과정에 충분히 반영하도록 하는 것이다. 일반적으로 사용자 프로파일 구축을 위한 학습은 학습 예제로 주어지는 도큐먼트의 속성(feature)을 기반으로 가중치를 부여하는 과정을 거쳐 이루어진다.

사용자에 의하여 제시된 자신의 선호도를 나타낼 수 있는 예제 도큐먼트들의 집합을 에이전트는 일종의 훈련집합(training set)으로 활용하여 학습하고 그 내용을 사용자 프로파일로 구축한다.

사용자 프로파일 모델(user profile model)로는 InfoFinder[9]에서의 불 서치 트리(boolean search tree) 모델과 키워드-벡터(keyword-vector) 모델 그리고 확률 모델등이 있으며 이중 키워드-벡터 모델이 웹과 같은 방대한 정보 공간에서 사용자가 원하는 정보를 추출하기 위하여 가장 널리 쓰이고 있다[10-14].

일반적으로 많은 정보 에이전트들이 도큐먼트의 키워드 벡터 표현을 위하여 도큐먼트로부터 추출된 모든 용어(term)들 -불용어(stop-word) 제외- 을 도큐먼트 속성(feature)으로 선택하고 속성의 가중치(weight)를 의하여 각 용어들이 나타난 문서와 전체 문서를 기준으로 한 빈도수에 기반한 TF*IDF 방식을 사용한다[14].

TF*IDF 가중치 $w_i = \log_2\left(\frac{N_i}{n_i}\right)$ (2)

w_i : i번째 용어의 가중치
N_i : i번째 용어의 빈도수(term frequency)
n_i : i번째 용어의 문서 빈도수(document frequency)

TF*IDF 가중치 식(2)에 의하면 도큐먼트 집합에 전체에 걸쳐 나타나는 용어들은 중요도가 낮고, 반대로 특정 도큐먼트에 나타나는 용어들은 상대적으로 중요도가 높다는 전제를 갖고 있다[20]. 따라서 임의의 도큐먼트 집합의 클러스터링 문제 있어서 도큐먼트 클러스터의 특성을 추출하기 의하여 일반적인 용어들을 제외시키는데 매우 유용하다[3]. 또한 TF*IDF 기법은 웹 도큐먼트의 HTML 태그(tag) 처리에도 적용되고 있다[14].

그러나 개인화된 지능적 정보 에이전트의 사용자 프로파일 구축을 위한 사용자 학습에 있어서 TF*IDF 가중치 방식은 다음과 같은 문제점이 있다. 먼저, 도큐먼트에 나타나는 용어 각각의 가중치만을 고려할 뿐 용어들간의 연관성(relation)을 파악 할 수 없다. 두 번째로 사용자가 자신의 선호도를 학습시킬 의도로 에이전트에게 제시하는 예제 도큐먼트들은 대부분 공통된 중심어를 포함하고 있다. 결과적으로 특정 주제 용어를 포함하는 양성(positive) 예제 도큐먼트 집합내의 주제어의 중요도가 상대적으로 낮게 나타나는 문제점을 갖고 있다.

따라서 용어들간의 관계성을 파악하고 그 가중치를 위한 새로운 기법들이 등장하고 있다. 정보의 듀얼리티(duality) 패턴을 이용한 방법[21], 휴리스틱(heuristic) 함수 기법[19], 단어 조합간의 코어커런스(co-occurrence) 빈도 통계[13] 그리고 데이터 마이닝 기법 등이 적용되고 있다[15].

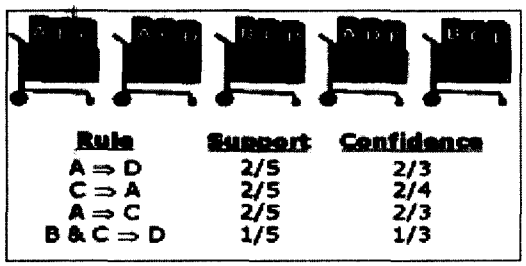
2.4 연관 규칙 탐색의 웹 마이닝 적용(Association Rule Exploration for Web Mining)

대부분의 키워드-기반 벡터 모델을 채용하고 있는 웹 검색 엔진들은 사용자에게 원하는 것을 정확히 입력하도록 요구하므로[10-12], 막연하나마 근접한 모든 정보를 찾고자 하는 사용자에게는 적합하지 않은 모델임 지적되었다[5].

본 논문에서는 짧은 사용자 질의(혹은 중심어)를 기준으로 연관된 도큐먼트 속성 객체들을 추출하여 사용자 프로파일을 확장시키기 위하여 웹 마이닝 (web mining)기법을 적용한다.

웹 마이닝 기법은 데이터 마이닝(data mining) 기법을 웹에 적용시킨 기술로, 웹으로부터 유용한 정보를 찾아내기 위한 기법이다. 주로 사용자에게 개인화된 웹 경험을 제공하기 위한 지식과 규칙을 생성하기 위하여 웹사이트 자체 혹은 사용자 로그 정보 등에 적용된다[22].

웹 마이닝은 그 적용 대상에 따라 웹 개인화 서비스를 위하여 웹 자체가 지닌 리소스를 분석하는 Web Content Mining과 사용자 접근 패턴을 파악하는 Web Usage Mining, 웹사이트와 웹 페이지의 하이퍼링크를 통하여 정보를 구조화시키는 Web Structure Mining등으로 분류할 수 있다[23].



(그림 1) 연관규칙 탐색 예제

개인화 웹서비스(personalized web service)를 제공하기 위하여 Web Usage Mining분야에 널리 적용되는 연관규칙 탐사 기법은 데이터 안에 존재하는 항목간의 종속관계를 찾아내는 기법을 말한다.

마케팅에서는 손님의 장바구니에 들어있는 품목간의 관계를 알아본다는 의미에서 장바구니분석(market basket analysis)이라고 한다[24]. 그림 1과 같은 장바구니 데이터베이스를 예로 들면, 품목 A와 품목 D의 지지도(support)=2/5 이고, 품목 A가 구매되었을 때 품목 D가 추가로 구매될 확률로서 품목 A의 품목D에 대한 신뢰도(confidence)=2/3 이다.

연관규칙 탐사 기법은 웹 마이닝에 적용되어 사용자 로그를 대상으로 연관 규칙을 탐사하여 마케팅에 응용되고 있으며 유사한 웹 도큐먼트들을 군집화 하기위한 특성 추출등에 적용되고 있다[25].

본 연구에서는 Web Content Mining 모델을 기반으로 연관규칙 탐색기법을 적용하여 웹 도큐먼트로부터 사용자의 선호도 혹은 관심과 연관 지식을 추출한다. 이를 위하여 연관규칙 탐사 기법을 웹 도큐먼트 환경에 적용한 연관 객체 추출 기법(Associate Object Extraction method)을 제시하고, 추출된 연관 객체의 가중치 부여(feature weighting)를 위하여 기존의 TF*IDF방식의 속성 가중치 기법을 보완한 연관 태그 블록 기반 가중치 기법을 제시한다.

3. 연관 객체 추출 기법(Associate Object Extraction method)

본 연구에서 지능적 정보 에이전트는 사용자가 제시한 예제 도큐먼트로부터 새로운 지식을 학습한다. 사용자의 정보요구와 관심 그리고 선호도를 반영하는 매우 중요한 의미를 지니고 있는 사용자질의를 에이전트의 사용자 학습을 위한 중심어로 삼아 그와 연관된 도큐먼트 객체(즉, 용어)를 추출하여 제시한다. 이를 위하여 데이터간의 연관성을 탐색해 내는 데이터 마이닝 기법인 연관 규칙 탐사 기법을 적용하여 연관객체를 추출한다. 웹 도큐먼트 기반 연관 지식 추출 알고리듬은 다음과 같다.

(표 1) 웹 도큐먼트 기반 연관 지식 추출 알고리듬

Alogrithm 1. Associate Object Extraction
1. Input wD : web document as a learning example O^Q : user query object 2. Document Preprocessing $oD \leftarrow$ *Content Parser*(wD) oD : Object_Document=a set of Tb Tb : Tag_Block-{Tag_{ID}, Tag_{Text}, $Tag_{Attribute}$} 3. Preprocessing Document Object $oT \leftarrow$ *Object Parser*(Tb) oT : Document_Object_Tuple=(Tag_{ID}, a set of O^C) O^C : Document_Content_Object 4. Explore Association Rule $aDB \leftarrow$ *Find Associates*(O^Q, oT) aDB : a set of aT aT : Associate_Object_Tuple=(O^Q, a set of O^C)

1. 에이전트는 사용자로부터 학습 예제 도큐먼트(web document as a learning example)와 사용자 질의 객체(user query object)를 입력받는다. 에이전트는 주어진 예제 도큐먼트로부터 사용자의 학습의도가 내포된 질의를 중심으로 연관 지식을 탐색하여 제시하는 것이 목적이다.
2. 연관 객체 탐색을 위하여 주어진 웹 도큐먼트를 파싱(ContentParser)하여 태그블록(tag block) 단위로 분해한다. 도큐먼트로부터 연관객체를 추출하기 위한 일반적인 도큐먼트 분해 방식으로는 문장-기반(sentence-based), 패시지-기반(passage-based) 방식이 있다[10][21]. 웹 도큐먼트 정보 컨텐츠의 표현과 배치(layout)를 위한 HTML형식상의 특징으로 인하여 보통 하나의 도큐먼트에 여러 토픽들을 포함하고 이들은 태그(tag)들에 의하여 구분된다[15]. 따라서 본 연구에서는 웹 도큐먼트로부터 연관 객체 탐

색을 위한 도큐먼트 분할 기준으로 태그 블록(tag block)을 제시한다.

3. 태그 블록에 나타나는 텍스트들에 불용어(stop_list) 처리, 어미 처리(stemming)등의 텍스트 전처리를 적용시킨 후 도큐먼트 객체 튜플(Document_Object_Tuple)을 생성한다.
4. 그리고 도큐먼트 객체 튜플(Document_Object_Tuple) 집합내에서 사용자 질의 객체와의 연관 객체를 추출(FindAssocaites)하여 연관 객체 튜플(Associate_Object_Tuple) 집합을 생성한다. 연관 객체 튜플 집합은 사용자 질의를 포함하고 있는 모든 태그 블록의 객체를 포함한다.

위 알고리듬의 결과로 생성된 연관 객체 튜플 집합은 웹 도큐먼트 상에서 사용자가 제시한 학습 중심어인 사용자 질의어와 같은 태그 블록내에 나타나는 용어들의 집합이다. 이는 용어들의 관계성을 발견해 내기 위하여 언어 집단 내 전체 용어 컬렉션을 분석하는 글로벌 컨택스트 기법에 비하여 상대적으로 적은 통계적 요구와 비용으로 효과적이고 안정적인 결과를 보이는 것으로 제시된 로컬 컨택스트 분석(local context analysis) 기법에 기반한다[8].

4. 태그 블록 기반 연관객체 가중치 기법 (Tag Block-based Associate Object Weight Method)

위 3절의 알고리듬에 의하여 추출된 연관 객체들은 사용자 중심어와 각 객체들 간의 신뢰도(confidence)값을 그 가중치로 부여받는다. 본 논문에서 제시한 연관 태그 블록 기반 가중치 기법은 사용자 중심어와 같은 블록에 나타난 객체들의 가중치를 사용자 중심어의 빈도수에 기반하여 조정하는 기법이다. 즉 중심어의 빈도가 높은 블록 내의 연관 객체들은 중심어의 빈도수가 낮은 블록 내에 나타난 객체들 보다 높은 연관 가중치를 받게 된다. 본 연구에서는 이를 위하여 물리학의 만유인력 모델을 적용한다. 만유인력이란 우주 공간에 있는 모든 물체사이에 작용하는 인력을 말한다. 공간상에 위치한 두 물체의 질량이 각각 M, n 라 할 때, 두 물체 사이에는 힘(F)이 작용한다(식(3)). 이때 G는 만유 인력 상수이고, r은 두 물체 사이의 거리이다.

$$F = G\frac{Mn}{r^2} \tag{3}$$

이와 같은 만유 인력 모델은 개체의 특성과 전체적인 분포 그리고 개체 상호간의 연관성을 고려한 자연스러운 군집화 기법에 적용되고 있다[26].

객체 연관성 공간상에서 객체들은 가중치가 증가함에 따라 사용자 프로파일에 합성되기 위한 순위의 기준인 부합치(correspondence weight)가 높아지게 된다.

(표 2) 연관 태그 블록 기반 가중치 알고리듬

Alogrithm 2. Tag Block-based Associate Object Weight (Universal Gravity Model)
1. Associate_Object_tuple aT^j=(O^Q, O^{C_1}, ..., O^{C_n})
2. Document_content_Object O^{C_i}: $\langle O_m^{C_i}, O_{cw}^{C_i} \rangle$
$O_m^{C_i}$: mass of O^{C_i}, tf
$O_{cw}^{C_i}$: weight of O^{C_i}
3. $F_j(O^Q, O^{C_i}) = G\frac{Q_m^Q \cdot O_m^{C_i}}{R(aT^j)^2}$ (4)
$R(aT^j)$: unit distance in aT^j
4. $O_{cw\_UG}^{C_i} = \sum_k D_k(O^Q, O^{C_i})$
$D_j(O^Q, O^{C_i}) = F_j(O^Q, O^{C_i}) \otimes f_{sig}$
f_{sig} : sigmoid function

1. 위 3절의 알고리듬에 의하여 추출된 연관 객체 튜플(aT^j)내의 객체(O^Q, O^C)들을 대상으로 가중치를 부여한다.
2. 각 도큐먼트 속성 객체들은, 각 개체들의 질량($O_m^{C_i}$) 으로 대표되는 태그 블럭내 빈도수에 의하여 각 개체들이 받는 힘에 따라 가중치

($O_{cw}^{C_i}$)가 부여된다.

3. 사용자 중심어 즉 중심 객체(O^Q)와 연관 객체 (O^C) 사이에 작용하는 힘을 측정하기 위하여 만유인력 모델 식(3)을 적용한다. 질량이 무거운 객체가 더 많은 힘(F)을 작용시키므로 중심 객체의 빈도가 높은 블록내의 연관 객체들의 중요도가 그만큼 증가하므로 에이전트의 학습에 있어 연관 지식들의 지역성(locality) 특징을 반영할 수 있다[8].
4. 연관 객체 튜플 단위로 사용자 중심어(O^Q)와 각 연관 객체(O^C)사이에 작용하는 힘을 기반으로, 객체의 가중치(O_{cw_UG})를 합산하여 부여한다.

5. 실험 및 결과

본 연구에서는 생명정보(bioinformatics) 도메인 분야에 위에 제시된 연관 지식 추출 알고리듬과 가중치 부여 기법을 적용하였다. 연관 지식 추출로 구축된 사용자 프로파일의 성능은 최소한의 키워드로 최대한의 의미를 설명하는 데 좌우된다. 따라서 중요도 가중치 순위상 상위에 연관 객체들을 얼마나 많이 랭크 시키는가가 객체 가중치 기법의 주요 역할이다.

실험에 사용된 사용자 중심어 "apoptosis"는 "세포사멸", 혹은 "세포자살"이라는 의미로 예제 도큐먼트에 "Cell Death", "Cell Suicide" 라는 연관 객체와 인접하여 자주 나타난다[7].

월드 와이드 웹과 같은 비구조적 정보들로 이루어진 데이터베이스를 D라 하고, 찾고자 하는 연관 타겟 객체 집합을 $R=r_1,\ldots,r_n$이라 하면, $R\subseteq D$ $R'\subseteq D$에 대하여, R'의 포함률(coverage), 정확률(precision), 오류율(error_rate)을 다음과 같이 정의한다.

표 3은 각 기법에 의한 가중치 순위를 기준으로 상위 10위(Top_10)의 용어들을 추출한 결과이다. "cell"과 같이 반전된 용어들은 "apoptosis"의

$$\text{포함율 } C(R,R')=\frac{|R'\bigcap R|}{|R|} \quad (5)$$

$$\text{정확률 } R(R,R')=\frac{|R'\bigcap R|}{|R'|} \quad (6)$$

$$\text{오류율 } E(R,R')=\frac{|R'-R|}{|R'|} \quad (7)$$

의미상 정의[1]에 나타나는 용어들이다. 반면, 밑줄 친(e.g. pombe) 용어들은 전체 용어들 중 사용자 중심어 "apoptosis"에 대한 연관 객체 튜플 집합에 나타나지 않는 용어들이다. 다시 말하면 "apoptosis"와 같은 태그 블록 내에 한번도 언급되지 않은 용어들이다

TF*IDF방식으로 부여된 객체 가중치 순위 결과(표 3의 TF*IDF열) "death", "suicide"그리고 "cell"은 "apoptosis" 연관되어 나타나는 의미있는 객체이지만 그 가중치 순위(order)가 하위로 나타남을 알 수 있다. 결과적으로 사용자 중심어 "apoptosis"와 관련된 용어가 아닌 전체 예제 도큐먼트중 가장 적은 도큐먼트상에 나타나는 "pombe"등이 상위에 나타났다. 그 이유는 TD*IDF방식은 도큐먼트 내에 나타나는 용어들의 빈도수와 전체 도큐먼트중 용어들이 나타나는 도큐먼트의 빈도수를 고려하여 가중치를 계산하므로 전체 도큐먼트들에 모두 나타나는 용어들은 그 특성이 약하고 반대로 특정 도큐먼트에만 나타나는 용어들은 상대적으로 중요성이 강하다는 전제를 갖기 때문이다[20].

1) Mechanism that allows cells to self-destruct when stimulated by the appropriate trigger. It may be initiated when a cell is no longer needed, when a cell becomes a threat to the organism's health, or for other reasons. The aberrant inhibition or initiation of apoptosis contributes to many disease processes, incl. cancer. Though embryologists had long been familiar with the process of programmed cell death, not until 1972 was the mechanism's broader significance recognized. Apoptosis is distinguished from necrosis, a form of cell death that results from injury., Britannica Concise, http://education.yahoo.com/search/be?lb=t&p=url%3Aa/apoptosis, accessed 2001/9/24.

(표 3) 가중치 순위 결과

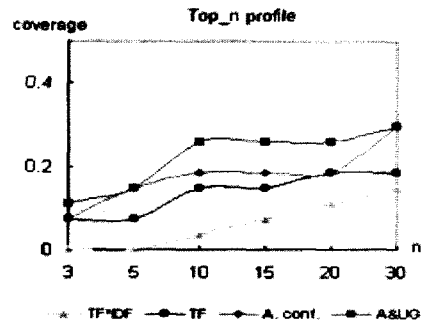
1	yeast	cell	apoptosi	apoptosi
2	pombe	apoptosis	cell	cell
3	cycle	yeast	image	cancer
4	fission	Image	cancer	image
5	fasl	page	trigger	trigger
6	caspase	suicide	suicide	process
7	apoptosis	pombe	signal	protein
8	division	fa	dy	suicide
9	institute	protein	infection	death
10	gene	death	apoptotic	dy

두 번째로, 질의 확장과 자동 개념학습 등을 위하여 같이 등장하는 용어(co-occurring term)들을 이용하여 연관 객체를 추출하는 기법들은 주로 예제 도큐먼트내의 용어들의 빈도수(TF)를 기준으로 추출한다[13]. 즉 가장 많이 등장하는 용어들을 상위로 랭크하는 기법이다. 그러나 이 기법 역시 속성 객체들간의 관계성을 반영할 수 없다는 한계점이 있다(표 3의 TF열).

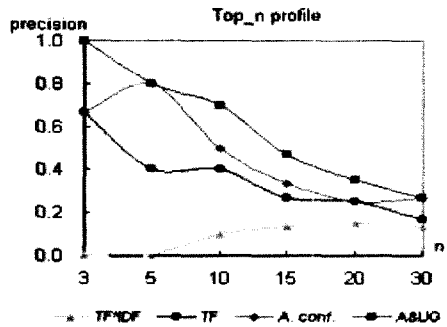
연관 규칙(association rule) 탐색 데이터 마이닝 기법을 텍스트 마이닝에 적용한 연구들에 의하면 도큐먼트 내의 용어들의 지지도(support)를 기준으로 속성을 추출하고, 신뢰도(confidence) 값을 가중치로 부여한다. 본 연구의 실험에서 사용자 중심어 "apoptosis"가 주어지는 에이전트 학습의 특성상 지지도는 고려하지 않는다. 단, 주어진 사용자 중심어 "apoptosis"에 대한 나머지 연관 객체들의 신뢰도(confidence)값을 가중치로 부여한 결과(표 3의 Association) 사용자 중심어와 매우 가까운 속성들이 선택되었다.

여기에 본 연구에서 제시한 만유인력 모델을 기반으로 한 연관객체의 관계성 가중치 기법에 의한 가중치 부여기법을 같이 적용한 결과(표 3의 A&UG열) 사용자가 제시한 주제어와 가까운 용어들이 보다 높은 가중치를 보여 결과적으로 사용자 프로파일에 통합되는 우선 순위 상위에 랭크됨을 알 수 있다.

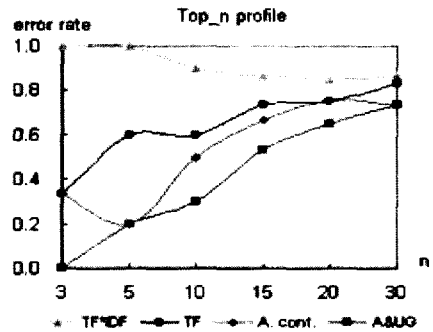
다음은 기존의 TF*IDF 가중치 기법(TF*IDF)과 TF가중치 기법(TF) 그리고 연관 규칙 탐색 기법만을 적용한 신뢰도(confidence)가중치 기법(A.conf.)을 적용한 결과와 본 연구에서 제시한 연관 지식 추출기법과 태그블록기반 가중치 기법(A&UG)을 적용한 결과중 상위 30위 속성에 대한 포함율(coverage), 정확율(precision) 그리고 오류율(error rate)의 비교 결과이다.



(그림 2) TF*IDF, TF, A.conf., A&UG : coverage



(그림 3) TF*IDF, TF, A.conf., A&UG : precision



(그림 4) TF*IDF, TF, A.conf., A&UG : error rate

실험결과 제시된 사용자 중심어 "apoptosis" 의 개념과 근접한 연관 지식을 효과적으로 추출함을 알 수 있다. 본 실험을 위하여 개인화된 지능적 정보 에이전트는 사용자 중심어 이외에는 어떠한 사전 지식도 부여되지 않았다. 제시된 예제 도큐먼트들은 다양한 정보원(yahoo, meta- crawler등)으로부터 수집된 웹 도큐먼트[27] 집합을 사용하였다. 또한 성능 평가 기준인 "apoptosis"의 사전적 정의[2]는 예제 도큐먼트에서 제외시켰다.

6. 결 론

지능적 정보 에이전트는 사용자 개개인의 성향에 맞게 정보를 가공하고 제공하는 기능을 수행하고 [17], 대량의 정보로부터 사용자를 보호하고, 새로운 지식을 탐색하여 사용자의 정보 가이드 역할을 수행하는 시스템이다[3]. 이러한 개인화 서비스 제공을 위하여 사용자 프로파일은 매우 중요한 역할을 한다. 그러므로 사용자 프로파일은 사용자의 관심과 선호도를 학습할 수 있어야 하며, 지능적으로 탐색된 관련 지식을 이용하여 확장되는 적응성(adaptability)이 그 핵심 속성이다.

본 논문에서는 사용자 로그 파일을 기반으로 접근(access)기록 분석을 통한 웹 마이닝 기법이 아닌 웹 컨텐츠 마이닝 기법으로서 적응적 사용자 프로파일 구축을 위한 웹 도큐먼트 기반 연관 지식 추출 기법을 제시하였다.

본 논문을 통하여 제시된 연관 지식 추출 기법은 사용자 중심어(즉, 질의어)와 관련 있는 도큐먼트 속성들을 추출하기 위하여 연관 규칙 탐사 기법을 적용하여, 같이 등장하고(co-occurred), 상호 관련 있는(inter-related) 속성들을 효과적으로 추출함을 알 수 있었다.

또한, 추출된 속성들의 가중치를 위하여 태그 블록 기반 가중치 기법은 사용자 중심어와 더불어 더 자주 언급되는 연관 도큐먼트 속성의 가중치를 증가시켜 그 관계성을 반영하도록 하였다.

위 두가지 기법을 모두 적용한 결과 기존의 단일 기법 적용 결과에 비하여 효과적인 에이전트의 프로파일 구축 학습 결과를 얻을 수 있었다.

이렇게 구축된 사용자 프로파일은 특정 도메인에 대한 정확한 사전 지식 없이도 에이전트 학습을 통하여 연관 지식을 학습하여, 자동 질의 확장이나 사용자에 대한 정보 가이드로서의 기능을 수행하는데 중요한 역할을 담당하며 방대한 비구조적 데이터베이스로부터의 새로운 지식 발견을 위하여 활용될 수 있을 것이다.

7. 향후연구

본 논문에서 제시된 사용자 중심 연관 객체 추출과 만유 인력 모델 기반 연관 객체 가중치 기법을 보다 다양한 생명정보(bioinformatics)분야로 확장 적용하여 알고리듬을 발전 및 개선시킬 필요성이 있다.

또한, 텍스트를 포함하여 이미지, 동영상 등의 멀티미디어 도큐먼트 객체들을 위한 연관 객체 추출을 위한 알고리듬과 시스템 구축이 필요하다.

또한 객체들간의 보다 세밀한 연관성 파악을 위하여 도큐먼트 내의 객체간 거리에 대한 표현과 측정을 위한 기법이 필요하다.

Acknowledgement

본 연구는 한국과학기술평가원 99 착수 숙명여자대학교 연구기반 확충사업, 정보시스템 보안을 위한 기반 및 응용 기술 연구과제의 지원으로 수행되었음.(과제번호 : 01-N6-02-01-A-02)

참 고 문 헌

[1] Kohrs, A. and B. Merialdo, Using category-based collaborative filtering in the active web-

2) Britannica Concise, http://education.yahoo.com/search/ be?lb=t&p=url%3Aa/apoptosis, accessed 2001/9/24.2)

museum, Institut EUROCOM-France 2000 IEEE, 2000.

[2] Bauer, M., Dengler, D., Paul, G., Instructible Information Agents for Web Mining, Proceedings of the 2000 international conference on Intelligent user interfaces, Pages 21 28, 2000.

[3] Iliopoulos, I., Enright, A. J., Ouzounis, C.A., TEXTQUEST: DOCUMENT CLUSTERING OF MEDLINE ABSTRACTS FOR CONCEPT DISCOVERY IN MOLECULAR BIOLOGY, PSB electronic proceeding 2001, http://psb.stanford.edu accessed 2001/9/25.

[4] Lin, C. and Mcleod, D. Temperament- based information filtering: A human factors approach to information recom-mendation, Computer Science department, University of Southern California, IEEE, 2000.

[5] Bigus, J. P. Constructing Intelligent Agent with Java, 1997.

[6] Luke,S., Spector, L., Rager, D., Hendler, J. Ontology-based Web Agents, Department of Computer Science, University of Marland, Proceedings of Autonomous Agents 97, USA, 1997.

[7] Xu, J., Croft, W.B., Improving the Effectiveness of Information retrieval with Local Context Analysis, ACM Transactions on Information Systems, Vol. 18, No. 1, pp. 79-112, January 2000.

[8] Krulwich, B., C. Burkey The InfoFinder Agent: Learning user Interests through Heuristic Phrase Extraction, AgentSoft Ltd., Andersen Consulting LLP, IEEE EXPERT, 1997.

[9] Armstrong, R., D. Freitag, T. Joachims and T. Mitchell,WebWatcher: A learning Apprentice for the World Wide Web, Proc. AAAI Spring Symp. on Information Gathering from Heterogeneous, Distributed Resources, Stanford, CA, 1995.

[10] Balabanovic, Marko, An Adaptive Web Page Recommendation Service, in ACM AGENTS 97, Proc. of the First International Conference on Autonomous Agents, Marina Del Rey, CA, 1997.

[11] Chen, Liren, and Kaita Sycara, WebMate : A Personal Agent for Browsing and Searching, in ACM AGENTS 98, Proc. of the International Conference on Autonomous Agents, Minneapolis, MN, 1998.

[12] de Lima, E. F., Pedersen, J. O., Phrase Recognition and Expansion for Short, Precision-biased Queries based on a Query Log, SIGIR99, pp. 145-152, 1999.

[13] Seo, Y., Zhang B., A Reinforcement Learning Agent for Personalized Information Filtring, Proceedings of the 2000 international conference on Intelligent user interfaces, pp. 248 251, 2000.

[14] Theeramunkong, T., Passage-Based Web Text Mining, Proceedings of the fifth international workshop on Information retrieval with Asian languages, pp. 205 206, 2000.

[15] Yang, J., Hong, K., Choi, J. "An Intelligent Collaborative Information Filtering Agent for Efficient Information Filtering", Proceedings of the 26th KISS Fall Conference, 1999.

[16] 정한혁, 이은석, 최중민, 한정현, 이준호, 지능형 전자 상거래를 위한 온토로지 서버구축과 개인 적응형 상품검색, 한국정보처리학회 논문지, 제7권, 제5호, 2000/5.

[17] Billsus, D., Pazzani, M. J., A Personal News Agent that Talks, Learns and Explains, Autonomous Agents99, ACM, pp. 268-275, 1999.

[18] Goecks, J., Shavlik, J., Learning Users Interests by Unobtrusively Observing Their Normal Behavior, Proceedings of the 2000 international conference on Intelligent user interfaces, pp. 129 132, 2000.

[19] Salton, G., A. Signhal, M. Mitra, C. Buckley Automatic Text Structuring and Summarization, Information Processing & management, v. 33(2), pp.193-207, 1997.

[20] Yi, J., Sundaresan, N., Mining the Web for Acronyms Using the Duality of Patterns and Relations, Proceedings of the second international workshop on Web Information and data management, pp. 48 52, 1999.

[21] Mobasher, B., Cooley, R., Srivastava, Automatic Personalization Based on Web Usage Mining, J. CACM, Vol. 43, No. 8, August, 2000.

[22] http://www.ciscorp.co.kr/ciscorp_web_ining2.htm, accessed 2001/8/29.

[23] http://home.pusan.ac.kr/~pnustat/info/Data Mining /2-1.htm, ac. 2001, 8/29.

[24] Cohen E. Datar M. Fujiwara S. Gionis A. Indyk P. Motwani R. Ullman JD. Yang C.,Finding interesting associations without support pruning, IEEE Transactions on Knowledge & Data Engineering , Vol. 13 No. 1, pp. 64-78, 2001

[25] Kim, E., Ko, J., Byun, H. Lee, Y. A Natural Clustering of Instances Based on Universial Gravity, Proceedinds of The 27th KISS Fall Conference, Vol. 27 No. 2, 2000.

◐ 저 자 소 개 ◑

문 현 정

1995년 숙명여자대학교 전산학과 졸업(학사)
1997년 숙명여자대학교 대학원 전산학과 졸업(석사)
2001년 현재 : 숙명여자대학교 대학원 컴퓨터과학과 박사과정
관심분야 : Intelligent Agent, Personalization, Multimedia
E-mail : hjmoon@cs.sookmyung.ac.kr

김 교 정

1972년 연세대학교 화학과 졸업(학사)
1983년 Clarkson Univ. 전산학 (석사)
1991년 Clarkson Univ. 전산학 (박사)
1986년 현재 : 숙명여자대학교 정보과학부 교수(멀티미디어 전공)
관심분야 : 인공지능, Multimedia
E-mail : kiochkim@sookmyung.ac.kr