# 사용자 중심 에이전트 학습을 위한 만유인력 모델기반 연관 객체 가중치 기법

문현정[0] 김교정
숙명여자대학교 정보과학부
(rosaline, kiochkim)@sookmyung.edu

# Universal Gravity Model-Based Associate Object Weighting for User-Centric Agent Learning

Hyunjeong Moon[0] Kio-Chung Kim
Dept. Information Science,
Sookmyung Women' s University

요 약

정보여과 에이전트는 자체의 적응성(adaptability)과 자율성(autonomy)을 특징으로 사용자의 선호도와 관심을 학습하여 사용자 프로파일을 지식베이스의 일부로 구축하는 기능을 수행한다. 이러한 사용자 프로파일은 사용자의 학습의도에 맞게 지식을 탐색하고 축적하는 적응성(adaptability)을 가져야 한다.

본 논문에서는 지능적 정보여과 에이전트가 사용자의 선호도와 관심을 학습하여 적응적인 사용자 프로파일을 구축하기 위한 기법으로서, 사용자가 제시한 학습예제로써의 웹 문서들로부터 사용자의 학습의도를 내포한 질의어를 중심으로 연관 지식을 탐색하여 추출하는 웹 도큐먼트 기반 사용자 중심 연관 객체 추출과 만유인력 모델을 기반으로 한 연관 객체 관계성 가중치 기법을 제시한다.

## 1. 서론

정보의 종류가 다양해지고 정보의 양이 증가할수록 사용자가 필요로 하는 정보를 찾기 위한 시간과 노력이 증가하는 정보과잉(information overload)상태가 발생하고 이를 해결하기 위하여 인텔리전트 정보 여과 에이전트(Intelligent Information Filtering Agent) 시스템이 활용되고 있다[1].

동적 학습 여과 모델을 기반으로 한 정보여과 에이전트는 사용자와 에이전트 간의 거듭된 인터렉션을 통하여 에이전트는 사용자의 선호도에 맞춰진 사용자 프로파일을 지식베이스의 일부로 구축하게 된다[2]. 이러한 사용자의 선호도에 따라 에이전트 자신의 지식내용을 변화시키는 에이전트의 적응성(adaptability)은 소프트웨어 에이전트의 기본적인 속성이며 동시에 개인화된(personalized) 정보여과 에이전트의 핵심적인 성격이다.

정보여과 에이전트의 사용자 프로파일(User Profile) 혹은 필터 (filter)를 구축하기 위하여 가장 많이 사용되는 기법은 내용기반 필터링(Content-based filtering)으로 TF*IDF 방식에 의한 가중치를 갖는 (키워드, 가중치)의 벡터공간 표현 모델이다[3]. 위 방법은 정보 여과 뿐 아니라, 전통적으로 문서에 대한 정보검색의 방법으로 널리 쓰이고 있다.

본 논문에서는 사용자의 학습의도를 중심으로 제시된 웹 도큐먼트들로부터 관계성 있는 용어들을 탐색하고 사용자 프로파일을 구축하기 위하여 만유인력 모델을 기반으로 한 연관 객체 관계성 가중치 기법을 제시한다.

## 2.관련연구

### 2.1 웹 마이닝

웹 마이닝은 데이터 마이닝 기법을 웹에 적용시킨 기술로, 웹 자체가 지닌 리소스를 분석하는 Web Content Mining 과 사용자 접근 패턴을 파악하는 Web Usage Mining, 웹사이트와 웹 페이지의 하이퍼 링크를 통하여 정보를 구조화 시키는 Web Structure Mining 등으로 분류할 수 있다[4]. 본 연구에서는 Web Content Mining 모델을 사용하여 웹 도큐먼트로부터 사용자의 선호도와 관심과 연관된 정보를 추출한다.

### 2.1 연관 규칙 데이터 마이닝 기법의 웹 마이닝 적용

본 논문에서는 Web Content Mining 방법으로 사용자가 제시하는 웹 도큐먼트로부터 사용자의 선호도와 관심과 연관된 용어를 추출하기 위하여 데이터 마이닝 기법 중에서 연관규칙 탐사 기법을 적용한다.

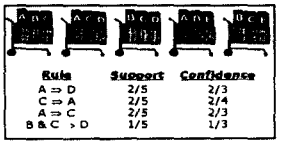


그림 1.연관규칙 탐색 예제

연관규칙 탐사란 데이터 안에 존재하는 항목간의 종속관계를 찾아내는 작업을 말한다. 마케팅에서는 손님의 장바구니에 들어있는 품목간의 관계를 알아본다는 의미에서 장바구니분석(market basket analysis)이라고 한다[5]. 그림 1 과 같은 장바구니 데이터베이스를 예로 들면, 품목 A 와 품목 D 의 지지도(Support) = 2/5 이고, 품목 A 가 구매 되었을 때 품목 D 가 추가로 구매

될 확률로서 품목 A 의 품목 D 에 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[6]

### 2.2 만유 인력 모델

만유인력 이란 우주공간에 있는 모든 물체사이에 작용하는 인력을 말한다. 공간상에 위치한 두 물체의 질량이 각각 M, n 라 할 때, 두 물체 사이에는 힘 $F=G\frac{Mn}{r^2}$ 이 작용한다. 이때 G는 만유 인력 상수 이고, r은 두 물체 사이의 거리이다.

이와 같은 만유 인력 모델은 개체의 특성과 전체적인 분포 그리고 개체 상호간의 연관성을 고려한 자연스러운 군집화 기법에 적용되고 있다[7].

본 연구에서는 웹 도큐먼트 학습예제로부터 사용자의 관심과 선호도를 기반으로 연관 용어들을 추출하고 관계성 가중치 부여를 위하여 만유 인력 모델을 적용한다.

## 3. 웹 도큐먼트 기반 사용자 중심 연관객체 추출 (User Centric Associate Object Extraction)

본 연구에서 에이전트는 인공지능 기계 학습 모델 중 예제 기반 학습 모델을 채용하여 사용자가 제시한 예제 도큐먼트로부터 새로운 지식을 학습한다. 사용자 질의는 사용자의 정보 요구와 관심 그리고 선호도를 반영하는 매우 중요한 의미를 지니고 있으므로 이를 기반으로 데이터간의 연관성을 탐색해 내는 데이터 마이닝 기법인 연관 규칙 탐사 기법을 적용하여 연관객체를 추출한다.

1. Input
   $wD$ : Web Document as a learning example
   $O^Q$ : User query object
2. Document preprocessing
   $oD \leftarrow ContentParser(wD)$
   $oD$ : Object_Document = a set of $Tb$
   $Tb$ : Tag_Block = { $Tag_{ID}$, $Tag_{Text}$, $Tag_{Attribute}$ }
3. Preprocessing Object Data
   $oT \leftarrow ObjectParser(Tb)$
   $oT$ : Document_Object_Tuple = { $Tag_{ID}$, a set of $O^C$ }
   $O^C$ : Document_Content_Object
3. Explore Association Rule
   $aDB \leftarrow FindAssociates(O^Q, oT)$
   $aDB$ : a set of $aT$
   $aT$ : Associate_Object_Tuple ={ $O^Q$, a set of $O^C$ }

그림 2. 웹 도큐먼트 기반 사용자 중심 연관 객체 추출 알고리듬

그림 2 의 알고리듬에 의하여 에이전트는 웹 도큐먼트 내의 연관 객체 추출을 위하여 사용자로부터 학습 예제 도큐먼트로(Web Document as a learning example)부터 사용자 질의 객체(User query object)를 입력 받는다. 연관 객체 탐색을 위하여 주어진 웹 도큐먼트를 파싱(ContentParser)하여 태그블록(Tag_Block) 단위로 분해한 후 도큐먼트 객체들의 튜플(Document_Object_Tuple)을 생성한다. 그리고 사용자 질의 객체와 연관 객체를 추출(FindAssocaites)하여 연관 객체 튜플(Associate_Object_Tuple)들의 집합을 생성한다.

## 4. 만유인력 모델을 기반으로 한 연관 객체 가중치 기법 (Universal Gravity model-based Associate object Weighting)

위 그림 2 의 알고리듬에 의하여 추출된 연관 객체들의 가중치를 부여하기 위하여 다음과 같은 연관 객체들의 만유 인력 공간 모델과 연관 객체 가중치 알고리듬을 제시한다.

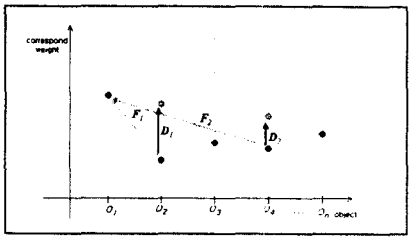


그림 3. 객체 연관성 공간 모델

1. Associate_Object_Tuple $aT^i=\{O^Q, O^{C_1}, \ldots, O^{C_n}\}$
2. Document_Content_Object $O^{C_i} : \langle O^{C_i}_{mass}, O^{C_i}_{cw} \rangle$
   $O^{C_i}_{mass}$ : mass of $O^{C_i}$, $tf$
   $O^{C_i}_{cw}$ : correspondence weight of $O^{C_i}$
3. $F_i(O^Q, O^{C_i}) = G\frac{O^Q_{mass} \cdot O^{C_i}_{mass}}{R(aT^i)^2}$
   $R(aT^i)$ : a unit distance in $aT^i$
4. $D_i(O^Q, O^{C_i}) = F_i(O^Q, O^{C_i}) \otimes f_{sig}$
   $f_{sig}$ : sigmoid function
5. $O^{C_i}_{cwNew} = O^{C_i}_{cwOld} + D_i(O^Q, O^{C_i})$

그림 4. 만유인력모델 기반 연관객체 가중치 부여 알고리듬

위 공간 상에서 객체들은 가중치가 증가함에 따라 사용자 프로파일에 합성되기 위한 순위의 기준인 부합치(correspondence weight)가 높아지게 된다. 위 3.1 에서 추출된 연관 객체 튜플 내의 객체들 간에는 서로 영향을 미치는 힘(F)이 작용하며, 각 개체들의 질량으로 대표되는 태그 블럭내 빈도수에 의하여 각 개체들이 받는 힘이 달라지게 된다. 즉 질량이 무거운 객체가 더 많은 힘을 작용하게 되므로 중심 객체의 빈도가 높은 블록 내의 객체들의 중요도가 그만큼 증가하므로 에이전트의 학습

에 있어 연관 지식들의 지역성(locality) 특징을 반영할 수 있다[8]

## 4. 실험 및 결과

본 논문에서 제시한 만유인력 모델을 기반으로 한 연관 객체의 관계성 가중치 기법을 생명공학분야를 도메인으로 하여 특정 주제어를 기반으로 검색된 웹 도큐먼트들을 에이전트의 학습예제로 제시하여 제시된 주제어 혹은 사용자 질의와 연관성 있는 객체들 중 그 가중치 순위에 따라서 에이전트의 지식 확장에 적용하였다[표 1].

표 1. TF*IDF vs. UGAW 가중치 기법 적용 결과

| tf | df | Object (TF*IDF) | rank | Object (UGAW) |
|---|---|---|---|---|
| 31 | 1 | yeast | 1 | apoptosis |
| 25 | 1 | pombe | 2 | cell |
| 17 | 1 | cycle | 3 | ccancer |
| 13 | 1 | fission | 4 | image |
| 12 | 1 | fasl | 5 | trigger |
| 11 | 1 | caspase | 6 | process |
| 55 | 6 | apoptosis | 7 | protein |
| 10 | 1 | institute | 8 | suicide |
| 10 | 1 | division | 9 | death |
| 14 | 2 | gene | 10 | dy |
| ... | ... | ... | ... | ... |
| 19 | 4 | death | 19 | ... |
| ... | ... | ... | ... | ... |
| 26 | 5 | suicide | 21 | ... |
| 174 | 9 | cell | 1048 | ... |

실험에 사용된 사용자 질의 "Apoptosis"는 "세포사멸", 혹은 "세포자살"이라는 의미로 예제 도큐먼트에 "Cell Death", "Cell Suicide" 라는 연관객체와 인접하여 자주 나타나는 것을 알 수 있다[9]. 즉 사용자가 에이전트의 사용자 프로파일구축을 위하여 학습예제를 제시하는 의도는 "Apoptosis"와 관련된 연관된 지식을 에이전트가 학습하여 추출된 지식을 이용하여 사용자 프로파일을 확장에 시키는데 그 목적이 있다. 따라서 사용자에 의하여 제시되는 학습예제는 특정 주제어를 공통적으로 포함하고 있으므로 학습 예제 도큐먼트 전체에 걸쳐 나타난다.

[표 1]의 결과에 의하면 TF*IDF 방식으로 부여된 객체 가중치 순위중 "death", "suicide" 그리고 "cell" 은 "apoptosis" 연관되어 나타나는 의미있는 객체이지만 그 가중치 순위(order)가 하위로 나타남을 알 수 있다. 본 연구에서 제시한 만유인력 모델을 기반으로 한 연관객체의 관계성 가중치 기법에 의한 가중치 부여 순위를 비교한 결과 사용자가 제시한 주제어와 가까운 용어들이 높은 가중치를 보였다. 그 이유는 TD*IDF 방식은 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[3].

## 5. 결론

본 논문에서는 지능적인 정보여과 에이전트의 개인화 서비스를 위한 적응적 사용자 프로파일의 구축을 위하여 사용자가 제시하는 학습 예제인 웹 도큐먼트로부터 사용자의 질의를 중심으로 연관된 도큐먼트 객체를 추출하기 위하여 연관 규칙 탐사기법을 적용하였으며 그 가중치 부여를 위하여 만유인력 모델에 근거한 연관 객체 가중치 기법을 제시하였다. 실험 결과 도큐먼트 내의 용어 가중치 부여를 위한 대표적인 기법인 TF* IDF 방식보다 학습 의도 주제어와 보다 가까운 용어들이 보다 높은 가중치를 보이므로 에이전트의 사용자 선호도 학습에 보다 적합한 것으로 나타났다.

앞으로 웹 도큐먼트상에서 동일 태그 블록내에 나타나는 객체들간의 거리와 빈도를 고려한 가중치 전략에 대한 연구가 필요하며, 웹 도큐먼트내의 텍스트 객체와 더불어 멀티미디어 객체들에 대한 연관성 가중치에 대한 연구가 필요하다

## 6. 참고 문헌


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

[2] Pannu, A. S., K. Sycare "A Learning Personal Agent for Text Filtering and Notification", *Proceedings of the International Conference of Knowledge Based Systems*, 1996.

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

[4]http://www.ciscorp.co.kr/ciscorp_web_mining2.htm, ac. 2001, 8/29.

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

[6] 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 , V.13 N.1 , 64-78 , 2001

[7] Kim, E., Ko, J., Byun, H. Lee, Y. "A Natural Clustering of Instances Based on Universial Gravity", Proceedings of The 27th KISS Fall Conference, V.27 N.2, 2000

[8] Wei Wang , Jiong Yang , Richard Muntz, STING+: An Approach to Active Spatial Data Mining,. Proceedings of the 15th International Conference on Data Engineering , 116-125 , 1999

[9] http://www.critpath.org/aric/library/img007.htm, ac.2001, 8/30.