온톨로지 기반의 계층적 개념 인덱싱을 이용한 사용자 관심사 학습

박지현° 김흥남 조근식
인하대학교 컴퓨터정보공학과
p9690@eslab.inha.ac.kr, nami@eslab.inha.ac.kr, gsjo@inha.ac.kr

Learning User Interest using Hierarchical Concept indexing based on Ontology

Ji-Hyun Park° Heung-Nam Kim Geun-Sik Jo
Department of Computer Science & Information Engineering, Inha University

요 약

인터넷의 급속한 성장과 더불어 사용자들은 인터넷을 통해 많은 정보를 얻을 수 있게 되었으며 최신 뉴스를 실시간으로 접근할 수 있게 되었다. 이에 따라 방대한 정보 속에 사용자 관심사에 맞는 정보를 효과적으로 검색하기 위한 여러 방법들이 연구되어 왔다. 하지만 기존의 많은 선행 연구들은 단어 빈도 기반의 키워드 벡터 모델을 이용하여 사용자의 관심사를 학습하고 있다. 이러한 키워드 벡터 모델은 사용자의 선호도를 명확하게 기술하지 못하고 키워드를 이용한 특징 벡터 (feature-vector)는 개념들 사이의 관계를 찾기 어려운 한계를 가지고 있다. 이를 개선하기 위해 본 논문에선 계층적 개념 인덱싱 (Hierarchical Concept Indexing)을 이용한 온톨로지 형태의 개인화된 사용자 프로파일을 만드는 방법을 제안한다. 생성된 사용자 프로파일에 개념 간의 유사도와 개념에 대한 사용자의 관심도를 고려하여 보다 개인의 선호도에 맞는 기사를 제공한다. 실험에서는 제안된 방법의 성능 평가를 위해서 기존의 키워드 벡터 모델의 학습 방법인 WebMate 시스템과 비교 분석하였다. 그 결과 제안하는 방법이 키워드 벡터를 이용한 학습 방법보다 향상된 성능을 보였다.

1. 서 론

인터넷 환경과 WWW (World Wide Web) 기술의 발달로, 현대 정보화 사회에서는 누구든지 웹 브라우저를 통하여 통제할 수 없을 정도로 수많은 정보와 뉴스들을 접근하고 이용할 수 있게 되었다. 이로 인해 수많은 정보와 뉴스 중에서 사용자가 원하는 정보만을 정확하고 빠르게 제공하는 것이 오늘날 정보 사회의 중요한 이슈가 되었다. 비록 검색엔진이 사용자들의 정보 검색을 도와주고 있지만, 매일 다양하게 변하고 새로 생성되는 정보와 뉴스 속에서 사용자가 원하고 필요로 하는 정보를 찾는 데에는 한계를 가지고 있다.

이러한 문제를 해결하기 위해서 사용자의 선호도에 따른 정보 검색 (Information Retrieval)과 정보 여과 (Information Filtering)에 대한 연구와 더불어 사용자의 성향을 효과적으로 학습하는 개인화 (Personalization)에 대한 연구도 활발히 이루어 지고 있다 [2][3][10]. 하지만, 개인화된 정보 여과 시스템의 핵심 요소인 사용자 프로파일 (User Profile)을 모델링하고 표현하는 많은 선행 연구들은 단어 빈도 기반의 키워드 벡터 모델을 이용한다. 이는 단어가 문맥에 따라 서로 다른 의미를 가지기 때문에 개념들 사이의 관계를 찾아내기 어려워 사용자의 관심사를 명확하게 기술하지 못한다는 문제점을 가지고 있다 [3].

따라서, 본 논문에서는 사용자의 프로파일을 계층적인 개념 트리 (Hierarchical Concept tree) 형태의 온톨로지 (Ontology)로 표현한다. 각각의 개념은 개념 인덱싱 (Concept Indexing) 기법을 이용하여 단어 (Term)와 단어의 가중치 (Weight)로 이루어진 벡터로 구성된다. 더불어 각 개념에 대한 사용자의 관심도를 고려하여 보다 개인화된 정보를 제공할 수 있는 학습 방법을 제안한다.

2. 관련 연구

2.1 개념 인덱싱 (Concept Indexing)

개념 인덱싱은 문서를 분류하는 방법의 하나로써, 키워드 대신에 개념을 사용하는 방법이다. 기존의 인덱스 가중치 부여 방법은 텍스트 내용 분석을 통해 단어 빈도수 (term frequency)를 이용하는 방법으로써 정확한 인덱스를 추출하기 어렵다는 단점을 가지고 있었다. 따라서 이를 보완하기 위해 단어뿐만 아니라 문서의 개념을 고려하는 인덱싱 방법이 연구되었다 [3]. 나아가 문서를 여러 개념으로 이루어진 복합 개념으로 간주하고 개념 벡터 모델을 이용해 더욱 정교한 문서분류를 위한 시도 및 키워드 추출과 같은 기존의 통계적 방법 대신, 상/하위 관계의 관련 개념으로 문서 인덱싱을 하는 방법이 연구되고 있다 [4][7].

2.2 개인화 (Personalization)

인터넷의 급속한 성장과 더불어 사용자의 성향을 효과적으로 학습하는 개인화 (Personalization)에 대한 연구도 활발히 이루어 지고 있다. WebMate [2]는 사용자가 웹을 효과적으로 탐색하고 검색하도록 도와주는 에이전트로 단어 빈도 기반의 키워드 벡터 모델로 사용자들 학습하여, 뉴스를 수집하고 자동으로 끌어와서 사용자에게 개인별 뉴스를 보내준다. 키워드 벡터 모델의 한계점을 개선하기 위해 PVA [1]는 사용자 프로파일을 계층화된 카테고리로 표현하였고, [3][10]에서는 다양하게 변화하는 사용자의 성향에 대한 연구를 수행하였다.

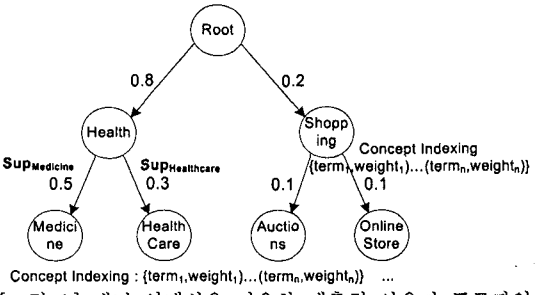
최근에는 웹 사용 마이닝 (Web Usage Mining)과 웹 내용

마이닝 (Web Content Mining)을 이용하여 보다 개인화된 서비스에 대한 연구도 활발히 이루어지고 있다 [5][6][11].

3. 사용자 관심 학습 (Learning User Interest)

3.1 사용자의 프로파일 표현 (User Profile Representation)

효과적인 사용자 프로파일을 모델링하고 표현하는 것은 개인화된 정보 여과 시스템의 핵심 요소이다 [10].
본 논문에서는 사용자의 관심사를 나타내는 사용자의 프로파일을 만들기 위해 개념과 개념 사이의 관계를 기술해 놓은 개념 계층 트리 구조의 온톨로지 형태로 표현한다.



[그림 1] 개념 인덱싱을 이용한 계층적 사용자 프로파일

[그림 1]은 사용자 프로파일의 개념 계층 구조 (Concept Hierarchy Tree)를 보여주고 있다. 사용자 프로파일에서 각 노드는 별개의 개념을 나타내고 방향성 있는 edge는 부모 노드와 자식 노드의 개념 관계를 나타낸다. 이러한 각각의 개념은 개념 인덱싱을 통해 단어와 단어의 가중치로 이루어진 특징 벡터로 나타낸다. 사용자가 관심 있는 학습 데이터들의 분포를 기반으로 각 개념의 지지도를 구한다.

3.2 계층적 개념 인덱싱 (Hierarchical Concept Indexing)

각 사용자 프로파일의 개념은 개념 인덱싱을 이용해서 특징 벡터로 표현한다. 개념 인덱싱은 키워드 대신 개념을 사용하는 방법으로 식 (1)를 이용해서 각 노드가 나타내는 각각의 개념에 속하는 단어들의 가중치를 구할 수 있다 [4].

$$E_{ik} = \log(f_{ik}+1.0)\times\left(1+\frac{1}{\log(N)}\sum_{j=1}^{N}\left[\frac{f_{ij}}{n_i}\log\frac{f_{ij}}{n_i}\right]\right) \quad (1)$$

여기서 E_{ik}는 개념 k에서 단어 i의 가중치 (weight)이고, f_{ik}는 개념 k에 단어 i가 나타나는 횟수, N은 개념의 총 개수, n_i는 단어 i를 포함하는 개념의 수를 나타낸다. E_{ik}의 값은 [0, 1]의 범위를 가지며 각 개념에 속한 단어 중요성을 의미한다. 즉, weight가 클수록 그 개념에서 단어의 중요성이 높음을 나타낸다.

개념 인덱싱 과정 후 각 개념들은 다음과 같은 특징 벡터로 정의될 수 있다.

$$CI_k = \{(term_1, weight_1),\cdots,(term_n, weight_n)\}$$

또한 다음과 같이 개념들 사이의 관계를 나타낼 수 있다.

$$\{(term_1, weight_1),\cdots,(term_n, weight_n)\} \xleftarrow{childof}$$
$$\{(term_1, weight_1),\cdots,(term_n, weight_n)\} \xleftarrow{childof} root$$

3.3 정보 여과 (Information Filtering)

본 논문에서는 사용자가 선호하는 문서를 여과하는데 두 가지를 측면을 고려한다.

첫 번째는, 사용자가 관심 있어하는 내용 기반으로, 이는 벡터 공간 모델에서 주로 사용되는 코사인 유사도를 확장한 방법을 이용한다. 새로운 문서 D_{new}는 TF/IDF를 이용해서 다음과 같은 특징 벡터로 표현 될 수 있다.

$$D_{new} = \{path,(term_1, weight_1),\cdots,(term_n, weight_n)\}$$

여기서 *path*는 문서가 속해있는 계층적 개념을 나타낸다.

해당 *path*의 상위 노드의 관심 정도가 훨씬 높음에도 불구하고 단순히 최하위 개념의 유사도가 낮아 추천이 안될 경우도 고려하기 위해서 최하위 개념의 유사도 뿐만 아니라 상위 개념의 유사도까지 고려한다. 만약 새로운 문서 D_{new}는 Root→B→E로 path가 정해져 있다고 가정한다면, 내용 기반의 유사도는 식 (2)과 같다.

$$CS(path) = \frac{\alpha\cdot sim(E, D_{new}) + \beta\cdot sim(B, D_{new})}{N} \quad (2)$$

$\alpha > \beta$, $\alpha+\beta=1$ 이고 $sim(V_i, V_j)$는 V_i와 V_j의 코사인 유사도 (Cosine Similarity), N은 개념의 개수를 나타낸다.

두 번째는 개념에 대한 사용자의 관심도를 고려한 것으로, 각 개념에 대한 지지도(support)는 식 (3)과 같이 정의되며, 이는 각 개념에 대한 사용자의 관심도를 의미한다.

$$\sup_k = \frac{n_k}{N} \quad (3)$$

여기서 k는 개념, N은 사용자의 관심사에 해당하는 문서의 총 개수, n_k는 N 중에서 개념 k에 속하는 문서의 개수이다.

최종적으로, 새로운 문서 D_{new}와 사용자의 프로파일의 유사도는 *path* 전체를 고려한 유사도 식 (2)와 개념에 대한 사용자의 관심도 식 (3)을 이용하여 식 (4)를 통해 측정한다.

$$Similarity(D_{new}, User) = CS(path)\times\sup_k \quad (4)$$

4. 실험 및 결과

4.1 실험 환경 및 데이터 집합

트레이닝 및 테스트에 사용된 데이터들은 RSS를 이용하여 야후 뉴스 (http://news.yahoo.com/rss)에서 2005년 4월부터 2005년 8월까지 수집한 기사를 사용하였으며, 데이터는 트레이닝 데이터 1013개, 테스트 데이터 412개로 구성된다.

사용자 10명이 총 1425개의 수집된 기사 중에서 관심 있는 기사만 선택하는 방법으로 피드백이 이루어졌다. 그 후 트레이닝 데이터로부터 야후 온톨로지 계층 구조에 따라 트리를 구성하고 개념 인덱싱으로 각각의 개념을 나타내는 각 사용자 프로파일을 구성하였다. 식 (2)에서 사용되는 개념 트리의 깊이 (depth)는 2로 하였으며, 테스트 데이터를 가지고 내용 기반의 개념 유사도와 개념에 대한 각 사용자의 지지도를 기반으로 사용자의 관심사에 맞게 뉴스를 추천하여 성능 평가하였다. 이때 새로운 문서의 *path*가 사용자 프로파일에 없는 경우 개념에 대한 지지도는 최하 값을 적용하였다.

4.2 실험 평가 방법

상위개념의 유사도와 카테고리에 대한 지지도를 추가해서 얼마나 정확하게 개인의 선호도에 맞는 기사를 제공하는지 성능을 평가하기 위해서 정확도 (Precision), 재현율 (Recall)과 F1-measure 측정식을 이용하였으며 각각의

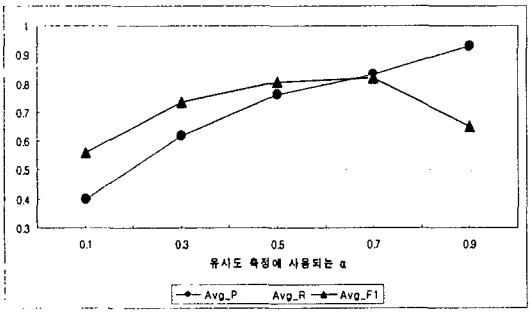
정의는 다음과 같다 [8].

$$F1-mesure=\frac{2\times recall\times precision}{recall+precision}$$

$$precision=\frac{N_c}{N_r},\ recall=\frac{N_c}{N}$$

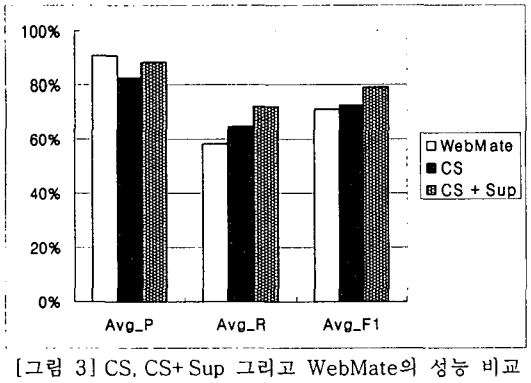
N_c는 정확하게 추천된 기사의 개수, N은 테스트 데이터 중 사용자가 선택한 기사의 개수이고 N_r은 사용자에게 추천된 기사의 개수이다.

4.3 실험 결과



[그림 2] α 값의 변화에 따른 평균 정확도, 재현율 및 F1-measure

[그림 2]는 사용자 프로파일과 새로운 문서 D_{new}의 유사도 측정에 이용되는 변수 α, β의 최적의 값을 측정하기 위해 평균 정확도 (Avg_P), 평균 재현율 (Avg_R) 그리고 평균 F1-measure의 값 (Avg_F1)의 변화를 나타낸 것이다. α의 수치가 높아질수록 평균 정확도는 올라갔으나 재현율은 낮아졌다. 하위 개념의 유사도에 높은 가중치를 주면 기존의 유사도에 적절한 비율로 상위 개념의 유사도를 반영할 수 있다. 반대로 하위 개념의 유사도에 낮은 가중치를 주면 상위 개념에 많은 가중치를 주게 되어 정확성이 떨어졌다. 실험 결과 하위 노드의 가중치 α 값이 0.7 일 때 성능이 가장 우수하였다.



[그림 3] CS, CS+Sup 그리고 WebMate의 성능 비교

[그림 3]은 제안하는 방법 중 개념 인덱싱의 유사도 만을 고려한 방법(CS), 개념 인덱싱의 유사도와 개념에 대한 사용자의 지지도를 함께 고려한 결과(CS+Sup) 그리고 키워드 벡터 모델인 *WebMate* [2]의 성능을 비교한 것이다.

실험 결과, WebMate가 정확도에서는 91%의 가장 좋은 성능을 보였으나, 재현율에서는 58%의 가장 낮은 성능을 보였다. 정확도와 재현율을 고려한 성능 평가 결과 개념 인덱싱의 유사도 방법 및 개념에 대한 사용자의 지지도를 함께 고려한 방법 (CS+Sup)이 개념 인덱싱의 유사도 만을 고려한 방법(CS) 보다 6.5%, *WebMate* 방법보다는 8.1% 높은 성능을 보였다.

5. 결론 및 향후 연구

본 논문에서는 사용자 개개인의 관심사에 알맞은 기사를 제공하기 위해 사용자 프로파일을 개념과 개념 사이의 관계를 기술하는 온톨로지를 이용한 개념적 계층 트리로 나타내었다. 그리고 사용자가 관심 있어 하는 내용 기반과 개념에 대한 사용자의 관심도를 고려하여 보다 개인화된 정보 서비스의 기틀을 마련하였다.

본 논문에서 제안한 학습 방법의 성능을 평가하기 위해서 기존의 키워드 벡터 모델의 학습 시스템과 비교 평가 결과 향상된 정보 여과 성능을 보였다. 하지만 본 논문에서는 사용자의 관심사의 변화를 고려하지 않고 있다. 그러므로 향후에는 빠르게 변하는 사용자의 관심사를 고려하는 연구가 필요하겠다.

6. 참고 문헌

[1] C. C. Chen and M. C. Chen, "PVA : A Self-Adaptive Personal View Agent,' Journal of Intelligent Information Systems, vol. 18, 2002.

[2] L. Chen and K. Sycara, "WebMate: Personal Agent for Browsing and Searching," In proc. of the 2nd Int. Conference on Autonomous Agents, 1998.

[3] M. E. Müller "Learning Scrutable User Models: Inducing Conceptual Descriptions," Journal of Kurzinformation, vol. 16, 2002

[4] S. S. Weng and C. K. Liu, "Using text classification and multiple concepts to answer e-mails" Expert Systems with Applications, vol. 26, 2004.

[5] B. Mobasher, H. Dai, T Luo, Y Sun and J Zhu "Integrating Web Usage and Content Mining" Lecture Notes In Computer Science; vol. 1875, 2000

[6] C. C. Aggarwal and P. S. Yu "An Automated System for Web Portal Personalization" In Proc. of the 28th VLDB Conference, Hong Kong, China, 2002.

[7] B. Y. Kang and S. J. Lee "Document indexing: a concept-based approach to term weight estimation," Information Processing and Management, vol. 41, 2005.

[8] Y. Yang, X. Liu. "A re-examination of text categorization methods," In Proc. of SIGIR-99, 1999

[9] C. Blake and W. Pratt "Better rules, few features: a semantic approach to selecting features from text." Proc. of the 2001 IEEE Int. Conference on Data Mining, 2001

[10] J. Wiley and Sons, "Learning user interest dynamics with a three-descriptor representation," Journal of the American Society for Information Science and Technology, vol. 52, 2001.

[11] B Mobasher "Web Usage Mining and Personalization" In Practical Handbook of Internet Computing, CRC Press, 2005.