# 데이터 마이닝과 집단 지성 기법을 활용한 소셜 콘텐츠 추천 방법에 대한 연구

강대현*, 박한샘*, 이정민*, 권경락*, 정인정*
*고려대학교 컴퓨터정보학과
e-mail : {internetkbs, park11232000, wjdals543, helpnara, chung}@korea.ac.kr

# A Study on Social Contents-Recommendation method using Data Mining and Collective Intelligence

Daehyun Kang*, Hansaem Park*, Jeungmin Lee*, Kyunglag Kwon*, In-Jeong Chung*
*Dept. of Computer and Information Science, Korea University

**요 약**

웹 기반 서비스의 발전과 스마트 기기의 보급으로 사용자들은 다양한 웹 서비스들을 이용할 수 있게 되었고, 소셜 웹과 같은 사람들 간의 관계를 형성함으로써 정보를 주고받는 서비스에 접근하여 자신만의 콘텐츠를 생성, 공유하기가 용이해졌다. 그러나 소셜 웹 사용자들이 증가하고 지식의 양이 늘어남에 따라, 방대한 양의 지식들 중 필요한 정보만을 효율적으로 창출해내고자 하는 연구 또한 시도되어 왔다. 그러나, 기존의 방법은 다수의 서비스 사용자들의 공통적인 관심사가 반영된 결과를 도출해내기에는 부족하다는 단점이 있었다. 그리하여, 본 논문에서는 집단 지성 알고리즘과 의사 결정 나무를 활용하여 소셜 웹을 이용하는 사용자들의 태그와 URL 정보를 토대로 트렌드를 분석, 콘텐츠를 추천하는 방법을 제안하고, 이를 통하여 다수 사용자들의 기호가 반영된 다양한 정보들을 소셜 웹 사용자들에게 제공해줄 수 있음을 보인다.

## 1. 서론

오늘날 웹 기반으로 제공되는 지식 서비스가 갈수록 증가하고, 온라인 소셜 네트워크와 같은 소셜 웹 서비스가 보급됨에 따라 수많은 사용자들이 이러한 웹 서비스와 웹 콘텐츠에 접근하기 수월한 환경이 마련되었다. 그리하여 소셜 네트워크, 소셜 커머스와 같은 소셜 웹 서비스들이 발전하게 되었으며, 이를 향유하는 사용자들은 갈수록 늘어나는 추세이다.[1] 이렇게 수많은 사용자들에 의해 생성, 공유되며 지속적으로 축적되는 유형의 지식들을 집단 지성[2]이라 한다. 이러한 집단 지성은 온라인 상에서 사용자 개개인의 창작물이나 지식이 다른 사용자들과 공유되기 용이한 커뮤니티 사이트 또는 위키피디아, P2P, 오픈 소스 등의 형태로 이용되고 있다.

소셜 웹을 이용하는 사용자들은 웹 서비스 상에 자신만의 콘텐츠를 생성하고 다른 사용자들과 공유함으로써 본인이 소유한 다양한 지식들을 이미지, 동영상, 태그와 같은 여러 형태로 만들어내고 있다. 그러나 소셜 웹 상에서 공유되고 생성되는 지식 자원의 양이 늘어나고, 자신이 원하는 정보를 더 빠르게 찾고 활용하고자 하는 사용자들이 늘어남에 따라, 소셜 데이터를 분석하여 수많은 지식들 중 개인 사용자에게 필요한 정보만을 빠르게 골라내고 제공해주기 위한 연구가 필요하게 되었다.

본 논문에서는 데이터 마이닝 기법 중 하나인 의사 결정 나무(Decision Tree)와 개미 군집 최적화(Ant Colony Optimization, ACO) 알고리즘을 활용하여 소셜 웹 사용자들의 태그 키워드들의 트렌드를 파악하고 사용자 콘텐츠 추천이 가능함을 보이고자 한다.

## 2. 관련 연구

### 2. 1. 집단 지성과 개미 군집 최적화 알고리즘

집단 지성이란 다수가 모여서 하나의 군집을 이루어서 활동하는 개체들이 집단적인 활동과 역할을 수행함으로써 얻어지는 결과물이나 그 과정 자체를 의미한다.[2] 집단 지성의 방법론으로는 개미[3], 벌[4], 바퀴벌레[5], 박쥐, 반딧불이 등과 같은 자연 속에서 같은 종끼리 군집을 이루어 행동하는 개체들의 습성에서 기인한 여러 알고리즘들이 존재한다[6].

본 논문에서 적용할 집단 지성 기법은 개미 군집 최적화 알고리즘이다. 이는 개미들이 목표점까지 가장 빠른 길을 협업적으로 찾아가는 과정을 정형화한 휴리스틱 알고리즘으로, 각 개미의 움직임에 따라 페로몬의 증발, 축적이 발생하며 이를 통해 최적의 길을 찾는다. 이러한 개미 군집 최적화 알고리즘의 원리는 온라인 소셜 웹 상의 다수의 사용자(개미)로부터 생성되고 축적된 콘텐츠(페로몬)를 기반으로 새로운 공통의 지식이나 경험을 추출, 분류하는 방법에 활용하기에 적합하다.

개미 군집 최적화 알고리즘에서는 어떤 꼭지점 $i$에

서 다른 꼭지점 $j$ 사이에 존재하는 페로몬의 양을 τ로 표기하며, $\rho$는 증발률, $m$은 전체 개미의 수를 나타내며, (식 1)과 같이 계산된다. 이후 반복적인 (식 1)의 계산을 통해 페로몬 양의 갱신이 이루어진다.

$$\boldsymbol{\tau_{ij} = (1-\rho) \times \tau_{ij} + \sum_{k=1}^{m} \Delta\tau_{ij}^{k}}$$

$$\Delta\tau_{ij}^{k} = \begin{cases} \dfrac{Q}{L_k} & k\text{번째 개미가 꼭지점 } i\text{에서 } j\text{로 이동한 경우} \\ 0 & \text{그 외 나머지 경우} \end{cases}$$

(식 1)

그리고 페로몬의 변화량 $\Delta\tau_{ij}^{k}$은 개미 $k$가 $i$에서 $j$로 이동했을 때 $i$와 $j$ 사이의 거리 $L_k$에서 임의의 상수 $Q$를 나눈 값으로 계산한다. 본 논문에서는 (식 1)을 아래 3 장에서 (식 2)로 변형하여 사용자들 사이의 콘텐츠에 대한 페로몬 값을 태그의 가중치에 따라 반복적으로 계산하고 의사 결정 나무를 구성하는 데 쓰인다.

2. 2. 의사 결정 나무

의사 결정 나무는 의사 결정 규칙(Decion Rule)을 나무 구조로 나타내어 분류와 예측을 위한 모델로서 주로 많이 활용되는 데이터 마이닝의 교사학습 방법론 중 하나이다. 이러한 의사 결정 나무의 구성에 쓰이는 알고리즘에는 ID3, C4.5[7], CN2[8], CART, CHAID 등이 있다. ID3 와 C4.5 는 다지분리 방식이며, 엔트로피(Entropy)와 정보 이득(Information Gain)을 각각 의사 결정 트리의 노드 분리 척도로 활용한다. 반면에 CART 와 CHAID 는 통계적 접근방식으로, 노드 분리 척도로서 지니 계수(Gini Index), 제곱 오차(Squared Error), F-measure 등의 방법을 활용한다. 의사 결정 나무는 적용 결과에 대해 명확하고 쉽게 이해할 수 있도록 도와주고, 정확도 또한 다른 분류 모델보다 우수한 편이며, 나무를 구성함에 있어서 입력 매개변수를 요구하지 않는다는 특성을 가진다. 그러나, 의사 결정 나무의 구축은 탐욕적 방법(Greedy Method)을 활용하기 때문에, 최적의 의사 결정 트리 구성을 위한 복잡도가 매우 크다는 단점이 있다.

본 논문에서는 소셜 웹 상에서 수집된 사용자들의 지식들을 활용하여 개미 군집 최적화 알고리즘에 따라 의사 결정 나무를 구성하고, 이를 통해 소셜 웹 사용자에게 필요한 정보를 제공해주고자 한다.

2. 3. 콘텐츠 추천 기법

콘텐츠 추천은 사용자의 선호도에 따라 적합한 대상을 제안하는 것을 말한다. 콘텐츠 추천은 사용자의 정보 탐색 효율을 높여주고 공급자, 수혜자 양쪽 모두에게 유익한 정보를 얻음으로써 생산성과 신뢰성이 증대된다는 장점이 있다.

콘텐츠를 추천할 대상을 선정하는 방법은 크게 두 가지로 분류된다. 콘텐츠를 추천받는 사용자의 개인 정보, 관심사와 같은 개인의 정보에 연관성이 높은 아이템들을 토대로 새 아이템을 추천하는 콘텐츠 기반 추천 기법[9]과, 콘텐츠를 추천해주고자 하는 사용자 주변의 다른 사용자들이 선호하는 아이템 정보를 토대로 새 아이템을 추천해주는 협업적 필터링 기법이 존재한다.[10] 콘텐츠 기반 추천 기법은 특정 사용자가 이전에 선호했던 개체들에 대한 정보를 토대로 새 아이템을 추천해 주는 방법으로, 개인의 프로필 정보나 관심사 정보가 많을수록 아이템의 다양성이 늘어나며, 아이템 간 유사도 측정을 위해 벡터 공간 모델을 이용하는 방법이 보편적이다.[11] 그러나, 콘텐츠를 제공받을 사용자에 대한 정보가 없거나, 지극히 적은 경우에는 추천의 정확도가 낮아진다는 단점이 존재한다. 반면, 협업적 필터링 방법은 특정 사용자가 속한 그룹의 관심사를 파악하여 콘텐츠를 추천해주는 방법으로, 특정 개인보다는 개인이 속한 그룹의 사용자들이 공통적으로 관심을 갖는 아이템에 중점을 둔다. 이 방법은 여러 사람의 정보를 토대로 아이템을 선별하기 때문에 그룹의 구성원이 많을수록 추천 아이템의 결과가 더 향상되고 종류 또한 다양해진다. 그러나, 이러한 그룹의 공통된 관심사를 얻기 위해 고려할 기준 사항이 많고, 개인보다 그룹에 중점을 둔 방법이기에 추천받은 콘텐츠에 대한 개인의 만족도가 다소 낮다는 단점이 있다. 이 외에도, 이러한 두 가지 방법 각각의 장단점을 보완하기 위하여 혼합한 하이브리드 추천 기법도 존재한다.[12, 13]

본 논문에서는 소셜 웹 상의 지식들을 URL 과 태그 형태로 수집하였고, 협업적 필터링 방법을 이용하여 다수 사용자들이 공통으로 선호하는 URL 링크와 태그를 분석하였다. 그리하여 이러한 분석된 내용과 유사한 정보를 얻고자 하는 다른 사용자에게 태그의 형태로 콘텐츠를 추천해주고자 한다.

## 3. 소셜 태그에 개미 군집 최적화 알고리즘을 적용한 의사 결정 나무 구성

본 논문에서는 소셜 태그를 분석하기 위하여 개미 군집 최적화 알고리즘을 의사 결정 나무에 적합하게 변형하여 이를 구성하는 방법을 활용한다.[14] (식 2)는 개미 군집 최적화 알고리즘의 변형된 식으로, 의사 결정 트리 구성 과정에서 노드의 페로몬 값을 계산하는 수식이다.

$$\boldsymbol{\tau_{m,m_L}(t+1) = (1-\rho) \times \tau_{m,m_L}(t) + \rho \times E} \qquad \text{(식 2)}$$

그리고 E는 의사 결정 나무의 평가 함수로서 다음 (식 3)과 같이 계산된다.

$$\boldsymbol{E(T) = \emptyset \cdot \omega(T) + \varphi \cdot \alpha(T,P)} \qquad \text{(식 3)}$$

(식 3)에서 $\omega(T)$는 의사 결정 나무 $T$의 크기를, $\alpha(T,P)$는 나무의 크기 $T$에서 실험 집단 $P$가 갖는 분류 정확도를, $\emptyset$와 $\varphi$는 나무의 크기와 분류 정확도 각각의 중요도를 결정하는 상수를 나타낸다. 본 논문에서는 $\omega(T)$는 선별된 태그의 종류를, $\alpha(T,P)$는 선별된 태그 종류 중 특정 태그가 갖는 비율을, $\emptyset$와 $\varphi$는 각

각 0.5 의 값을 갖도록 설정하였다.

## 4. 실험 및 결과

### 4.1. 데이터 수집

소셜 웹 사이트 중 하나인 Delicious[1] 에서는 URL 과 태그의 형태로 온라인 상에서 북마크(bookmark)를 저장, 공유함으로써 사용자 간 정보 공유가 이루어진다. 이 때, 한 URL 에서 검색된 태그 개수에 따라 해당 단어에 대한 가중치가 주어지고, 이러한 단어들이 한 URL 에 여러 개 존재한다. 다음 (그림 1)은 1033 명의 각기 다른 사용자가 저장한 하나의 고유 URL 과 이에 속한 여러 태그들이 웹 서비스 상에서 보여지는 한 예이다.


(그림 1) 실험에 사용된 데이터 중 하나의 고유 URL 과 이에 속한 여러 태그들의 예

실험에 활용하기 위해 Delicious 에서 1,123 개의 고유 URL 과 고유 URL 들에 속한 각기 다른 20,244 개의 태그를 xml 서식으로 수집하였고, 그 중에서 사용자들이 가장 많이 언급하고, 링크를 저장한 최상위 수준에 해당하는 URL 과 태그를 선별하여 알고리즘을 적용, 실험을 진행하였다.

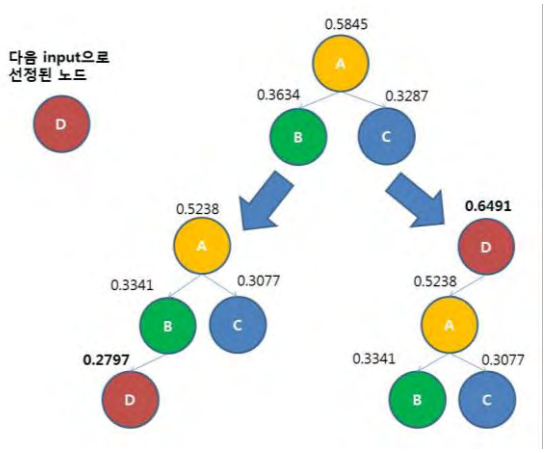
### 4.2. 의사 결정 나무 구성 및 추천 콘텐츠 선정

본 논문에서 개미 군집 최적화 알고리즘에 의해 실험에 활용된 태그들은 의사 결정 나무의 노드로 구성되며, 각각의 태그에 대한 페로몬 값이 지속적으로 갱신됨에 따라 의사 결정 나무의 전체 구조 또한 변화한다. 이 때, 어느 특정 노드가 의사 결정 나무의 다음 노드로 적합한지의 여부는 다음 (식 4)에 따라 확률적으로 결정된다.[14]

$$p_{i,j} = \frac{\tau_{m,m_{L(i,j)}}(t)^{\alpha} \cdot n_{i,j}^{\beta}}{\sum_i^a \sum_j^{b_i} \tau_{m,m_{L(i,j)}}(t)^{\alpha} \cdot n_{i,j}^{\beta}} \quad \text{(식 4)}$$

(식 4)에서, $\tau_{m,m_{L(i,j)}}(t)$는 시간 $t$일 때 노드 $m$과 노드 $m_L$ 사이의 페로몬 양을, $n_{i,j}$는 속성 $i$와 $j$를 검사하는 휴리스틱 임계값(threshold)을, $\alpha$와 $\beta$는 상관관계를 나타내기 위한 계수를 의미한다. 그리하여, 시간 $t$일 때의 특정 노드가 가진 페로몬의 양을 전체 페로몬의 양으로 나누어서 더 높은 값을 가진 노드가 의사 결정 나무의 다음 노드로 선정되고, 시간의 흐름에 따라 지속적으로 의사 결정 나무가 변화해나가면서 최종 단계까지 진화해간다. (그림 2)는 이러한 의사 결정 나무가 생성된 이후 성장해나가는 초기 과정을 보여준다.

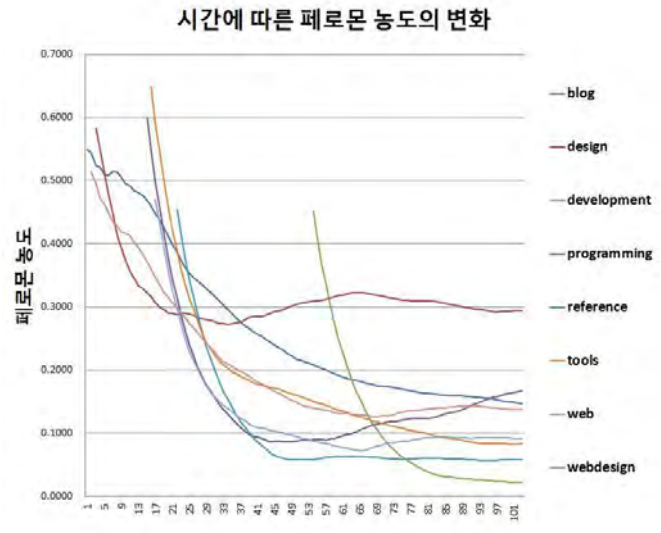
[1] www.delicious.com/




(그림 2) 의사 결정 나무의 초기 구성 단계

(그림 2)에서 확인할 수 있듯이, 각 노드는 현재 의사 구성 나무 상태에서의 자체 페로몬 농도를 지니며, 새로운 노드를 선별하는 과정에서 페로몬 양이 갱신됨에 따라 각각의 노드 쌍의 이전 페로몬 양을 (식 2)에 따라 변화시킨다.

본 논문에서는 (식 2)와 (식 3)을 활용하여 수집한 URL 과 태그 중 각기 다른 96 개의 URL 과 103 개의 태그 키워드 분석을 통해 페로몬 농도를 지속적으로 갱신하고 의사 결정 나무를 구성하였다. 이 때, 시간에 따른 페로몬 농도의 변화를 그래프로 나타낸 결과는 다음의 (그림 3)과 같다.

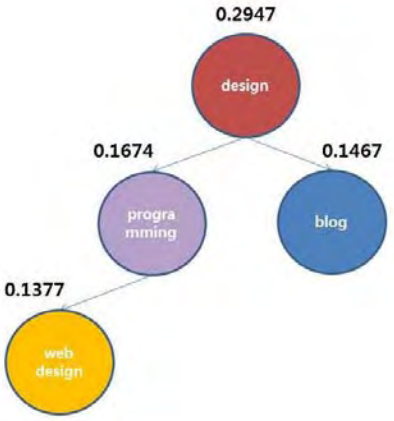



(그림 3) 의사 결정 나무가 구성될 때 시간에 따른 각각의 태그들이 가진 페로몬의 변화량

(그림 3)에서는 실험에 활용된 각각의 태그들이 갖는 페로몬 농도의 시간에 따른 변화량을 보여준다. 태그의 페로몬 농도는 태그가 처음 언급된 시점부터 줄곧 증발되어 감소하지만, 이후 사용자들의 해당 태그의 언급이 지속적으로 발생하면 페로몬의 농도의 증발이 적어지며 일정한 수준을 유지하거나 또는 더 축적된다는 사실을 알 수 있다. 또한, 어느 특정한 시점에서의 전체 태그들의 페로몬 양을 통해 특정 시점에서의 소셜 웹 사용자들이 해당 태그에 대해 갖는

선호도도 파악할 수 있다.

다음 (그림 4)에서는 태그 분석을 통해 형성된 의사 결정 나무의 한 사례를 보여준다.




(그림 4) 사용자 태그 분석을 통해 구성된 의사 결정 나무의 예

(그림 4)에서는 페로몬 값이 가장 높은 노드가 의사 결정 나무의 뿌리 노드이며, 그 하위 노드로 구성된 각기 다른 태그들이 있음을 알 수 있다. 이렇듯 페로몬 농도에 따라 최적화된 의사 결정 나무로 구성된 태그를 사용자가 언급할 경우, 페로몬 농도 별 순위와 함께 그 태그와 연관된 다른 태그로서 사용자에게 정보 제공이 가능하다.

## 5. 결론

온라인 상에서 웹 서비스의 형태로 사용자들에 의해 생산, 공유, 소비되는 콘텐츠들이 늘어나고, 이러한 콘텐츠를 향유하는 사용자 또한 갈수록 늘어나고 있다. 그리하여 보다 양질의 콘텐츠를 빠르고 효율적으로 확보하기 위한 노력 또한 지속적으로 요구되는 추세이다. 그리하여 데이터 마이닝, 집단 지성 등 여러가지 분석 기법들을 통해 수많은 양의 정보를 빠르게 처리하기 위한 여러 연구들이 시도되고 있다.

본 논문에서는 집단 지성 알고리즘 중 하나인 개미 군집 최적화 알고리즘과 데이터 마이닝 기법 중 하나인 의사 결정 나무를 활용하여 소셜 웹 사용자들의 트렌드를 파악하고 관련성이 높은 콘텐츠를 추천해주는 방법을 제안하였다. 소셜 웹 상에서 URL 과 태그 형태로 존재하는 개개인 사용자들의 관심사 정보를 토대로 주변 사용자들의 공통된 정보를 얻고, 새로운 정보를 얻고자 하는 개인 사용자들에게 콘텐츠를 제공해줄 수 있다. 향후 연구로는 제안한 방법을 이용하여 소셜 웹 사용자로부터 특정한 조건이 주어졌을 때, 본 논문에서 활용한 방법을 적용하여 타당성있는 콘텐츠 추천이 가능한지에 대한 검증과 더 큰 데이터 셋에 대한 적용이 필요하다. 이를 통해 태그와 같은 텍스트로 이루어진 소셜 웹 사용자들의 관심사들을 응집시켜 새로운 태그에 대하여 연관성이 더 높은 결과를 보여줄 수 있을 것으로 기대된다.

## 사사

본 논문은 교육 과학 기술부의 재원으로 한국 연구재단의 지원을 받아 수행된 BK21 플러스 사업의 연구 결과임 (No. T1300572)

## 참고문헌

[1] T. Gruber, "Collective knowledge systems: Where the social web meets the semantic web," *Web semantics: science, services and agents on the World Wide Web*, vol. 6, pp. 4-13, 2008.

[2] E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm intelligence*: Oxford, 1999.

[3] M. Dorigo and C. Blum, "Ant colony optimization theory: A survey," *Theoretical computer science*, vol. 344, pp. 243-278, 2005.

[4] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," *Journal of global optimization*, vol. 39, pp. 459-471, 2007.

[5] L. Cheng, Z. B. Wang, Y. H. Song, and A. H. Guo, "Cockroach swarm optimization algorithm for TSP," in *Advanced Engineering Forum*, 2011, pp. 226-229.

[6] M. Sajwan, K. Acharya, and S. Bhargava, "Swarm Intelligence Based Optimization for Web Usage Mining in Recommender System," *International Journal of Computer Applications Technology and Research*, vol. 3, pp. 119-124.

[7] J. R. Quinlan, *C4. 5: programs for machine learning* vol. 1: Morgan kaufmann, 1993.

[8] P. Clark and T. Niblett, "The CN2 induction algorithm," *Machine learning*, vol. 3, pp. 261-283, 1989.

[9] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, pp. 66-72, 1997.

[10] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: an open architecture for collaborative filtering of netnews," in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 1994, pp. 175-186.

[11] J.-S. Sohn, U.-B. Bae, and I.-J. Chung, "Contents Recommendation Method Using Social Network Analysis," *Wireless Personal Communications*, vol. 73, pp. 1529-1546, 2013/12/01 2013.

[12] Y.-Y. Shih and D.-R. Liu, "Hybrid recommendation approaches: collaborative filtering via valuable content information," in *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, 2005, pp. 217b-217b.

[13] M. Nilashi, O. B. Ibrahim, and N. Ithnin, "Hybrid recommendation approaches for multi-criteria collaborative filtering," *Expert Systems with Applications*, vol. 41, pp. 3879-3900, 2014.

[14] U. Boryczka and J. Kozak, "Ant colony decision trees–A new method for constructing decision trees based on ant colony optimization," in *Computational Collective Intelligence. Technologies and Applications*, ed: Springer, 2010, pp. 373-382.